UCLA-ENG-7166
October 1971

ARPA CONTRACT NO. DAHC-15-69-C-0285

ANALYSIS OF A CONTINUUM OF PROCESSOR-SHARING MODELS FOR TIME-SHARED COMPUTER SYSTEMS

Jiunn Hsu

COMPUTER SYSTEMS
MODELING AND ANALYSIS GROUP

COMPUTER SCIENCE DEPARTMENT

School of Engineering and Applied Science
University of California
Los Angeles

R
150

MISSING PAGE
NUMBERS ARE BLANK
AND WERE NOT
FILMED

COMPUTER SYSTEMS MODELING AND ANALYSIS GROUP
REPORT SERIES

1. Cole, G.D., "Computer Network Measurements: Techniques and Experiments," October 1971, UCLA-ENG-7165 (ARPA).
2. Hsu, J., "Analysis of a Continuum of Processor-Sharing Models for Time-Shared Computer Systems," October 1971, UCLA-ENG-7166 (ARPA).

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency of the U.S. Government.

published by

REPORTS GROUP
SCHOOL OF ENGINEERING AND APPLIED SCIENCE
UNIVERSITY OF CALIFORNIA, LOS ANGELES 90024

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)

University of California
School of Engineering and Applied Science
Los Angeles, California 90024

2a. REPORT SECURITY CLASSIFICATION

2b. GROUP

3. REPORT TITLE

ANALYSIS OF A CONTINUUM OF PROCESSOR-SHARING MODELS FOR TIME-SHARED COMPUTER SYSTEMS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (First name, middle initial, last name)

Jiunn Hsu

6. REPORT DATE: October 1971

7a. TOTAL NO. OF PAGES: 151

7b. NO. OF REFS: 47

8a. CONTRACT OR GRANT NO.: DAHC-15-69-C-0285

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S): UCLA-ENG-7167

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

Processor-sharing models of time-shared computer systems are defined and some new results are presented. The major emphasis of this research is on the modeling and analysis of new models of time-shared computer systems and on the finding of some fundamental properties which apply to the average number of customers in the system and the average response time functions for a large class of time-shared computer systems.

The family of selfish scheduling algorithms is defined and the Laplace transform of the response time functions are obtained. The selfish round robin (SRR) and the selfish foreground background (SFB) systems are given as two illustrative examples.

A family of scheduling algorithms whose performance ranges between that of the RR system and the FB system is constructed. A weighting function g(t) is given to define the scheduling algorithm such that a customer's rate of attaining service depends on how much service time t he has already got. The average response time function for this family of systems is obtained.

A simple relationship between the time-dependent average number of customers in the system and the average response time functions for a large class of M/M/1 systems is formulated. The result shows that the behavior of one customer can strongly influence the total number of customers in the system.

Finally, some fundamental properties are established which apply to the average response time functions for all time-shared computer systems. Among them, tight upper and lower bounds on the average response time are obtained.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Computers						
Time-Sharing						
Queueing						
Priority Queueing						
Analytical Time-Sharing Models						
Response Time Analysis						

UCLA-ENG-7166

ANALYSIS OF A CONTINUUM OF PROCESSOR-SHARING MODELS FOR TIME-SHARED COMPUTER SYSTEMS

by

Jiunn Hsu

Supported by

ADVANCED RESEARCH PROJECTS AGENCY
Computer Network Research
ARPA Order No. 1380, Program Code No. 9D30
Contract No. DAHC-15-69-C-0285

Leonard Kleinrock, Principal Investigator

Computer Systems Modeling and Analysis Group

Computer Science Department
School of Engineering and Applied Science
University of California
Los Angeles, California 90024

October 1971

PREFACE

The research described in this report, "Analysis of a Continuum of Processor-Sharing Models for Time-shared Computer Systems," by Jiunn Hsu, is part of a continuing investigation of Computer Network Research, sponsored by the Advanced Research Projects Agency (ARPA), Department of Defense Contract DAHC-15-69-C-0285, under the direction of L. Kleinrock, Principal Investigator, and D. Estrin, M. Melkanoff, and R. Muntz, Co-Principal Investigators, in the Computer Science Department of the School of Engineering and Applied Science, University of California, Los Angeles. This project was also partially sponsored by an IBM Fellowship.

This report was the basis of a Ph.D. dissertation (June 1971) submitted by the author under the chairmanship of Leonard Kleinrock.

ACKNOWLEDGMENTS

I wish to express sincere thanks to Professor Leonard Kleinrock for his guidance, patience and friendship during the preparation of this dissertation. I should also like to thank Professor R. R. Muntz for his invaluable suggestions and ideas. Appreciation also goes to Professors W. W. Chu, D. G. Cantor and L. A. Lippman for serving on the doctoral committee.

A debt of gratitude is owed to the International Business Machines Corporation for the granting of an IBM Fellowship.

This work was partially supported by the Advanced Research Projects Agency of the Department of Defense, Contract Number DAHC-15-69-C-0285.

I also wish to thank Mr. Tom T. Tang of the National Cash Register Company for his encouragement, Mrs. Charlotte B. LaRoche of the U.C.L.A. Computer Science Department for her assistance in preparing this dissertation.

Without the support and encouragement of my wife, none of this would have been possible.

ABSTRACT

Processor-sharing models of time-shared computer systems are defined and some new results are presented. The major emphasis of this research is on the modeling and analysis of new models of time-shared computer systems and on the finding of some fundamental properties which apply to the average number of customers in the system and the average response time functions for a large class of time-shared computer systems.

The family of selfish scheduling algorithms is defined and the Laplace transform of the response time functions are obtained. The selfish round robin (SRR) and the selfish foreground background (SFB) systems are given as two illustrative examples.

A family of scheduling algorithms whose performance ranges between that of the RR system and the FB system is constructed. A weighting function g(t) is given to define the scheduling algorithm such that a customer's rate of attaining service depends on how much service time t he has already got. The average response time function for this family of systems is obtained.

A simple relationship between the time-dependent average number of customers in the system and the average response time functions for a large class of M/M/1 systems is formulated. The result shows that the behavior of one customer can strongly influence the total number of customers in the system.

Finally, some fundamental properties are established which apply to the average response time functions for all time-shared computer systems. Among them, tight upper and lower bounds on the average response time are obtained.

CONTENTS

Page

FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Time-Sharing Computer Systems

The value of time-shared processing systems as a means of providing a processor to many users concurrently is well-established. The rationale for most time-sharing systems is to provide fast service for customers with short, highly interactive programs in order to facilitate debugging, to encourage experimentation with improvisation of computing methods, and to support general interactive computations. In the typical operation of such a system, the users communicate with the computer by means of teletype or similar I/O devices. As each user makes a request for computer processing, he in effect enters a queue whose members are served in a way determined by the specific scheduling algorithm being used. Here we define a scheduling algorithm as a set of decision rules determining which user will next be serviced and how long he will be given use of processing facilities. Thus each program in turn is transferred into memory, operated upon and transferred out. It is obvious, unless this swapping of programs can be done at no cost in time or the memory is large enough so that no swapping of programs is needed, that this mode of operation is less efficient than the batch system where each request is run to completion. The technique of time-sharing, however, results in faster average response time for the user with short request. This fast response makes it appear, to such a user, that he is the only one using the computer.

The effectiveness of the time-sharing systems depends in large part on the efficiency with which the resources are allocated to the individual users. Thus, considerable attention has been focused on the time and space scheduling problems of time-sharing systems and many analytical results have been obtained [1] since the appearance of the first applied paper published in 1964 [2]. In most of the results, only one resource (the CPU) is to be shared, where it is assumed that the size of the main memory is infinite. There are two reasons for this assumption: in the past, most of the tools used to analyze time-sharing systems have been drawn from queueing theory, and it is very difficult to analyze a system with two resources (thus two queueing structures) which are not independent of each other. The second reason is that modeling of program behavior and peripheral devices is very difficult. Of course, the assumption of infinite memory somewhat weakens our models. However, the CPU is one of the most important resources in the computer system, as long as the size of the main memory is adequate (so that the system is not memory bound). The analysis of single resource systems gives us a good feeling of how time-sharing systems behave. In this dissertation, we concentrate on the single resource case. We analyze a class of such systems and also give some general behavior constraints.

1.2 An Existing Time-Sharing Computer System

In this section we describe an existing time-sharing computer system. We choose the Model 67 of the IBM System/360 as our example. The following description is quoted from Gibson [3].

The basic architecture of the IBM System/360 makes it well suited to processing in a multiprogramming and multiprocessing environment. The Model 67 extends this basic architecture to provide the additional capabilities ot an advanced time-sharing system.

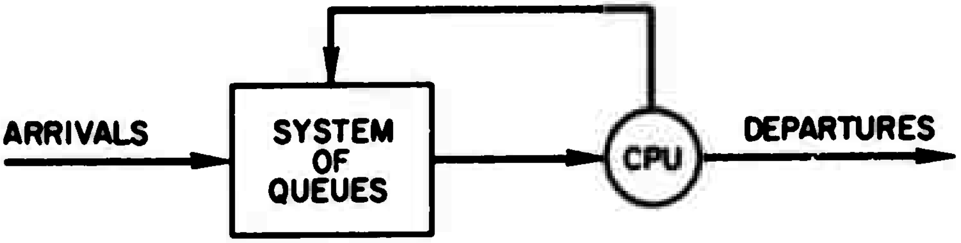
The Model 67 incorporates multiprogramming, multiprocessing, and multiaccess capabilities. Multiaccess allows several users at remote consoles to communicate directly with the system and to present a number of applications ranging from conversational compiling to desk calculator functions. Multiprogramming is defined as the ability to have several active programs reside in core simultaneously. As soon as one job is finished, or is held up by an I/O request, or has depleted its time allowance, the next task can begin immediately.

The dynamic relocation feature built into the hardware facilitates multiprogramming; peripheral operations will now be just like any other tasks in the memory. Even without the multiaccess capability, multiprogramming provides much more efficient utilization of the computer's resources than in a stacked job operation. For the first time, a central processing unit is a resource that can be allocated. With multiaccessing, where some of the jobs in core belong to remote terminals, the multiprogramming capability is further enhanced as this enables the rapid switching between jobs, or "time-slicing."

The Model 67 enables each processor of a multiprocessor system to operate as a single processor with its own I/O subsystem, or jointly with other processors in a symmetric multiprocessing configuration.

1.3 The Mathematical Model

Figure 1.1 shows a general feedback queueing model where the CPU is being shared. Incoming jobs are queued and scheduled for service in some way. At its scheduled service time, each job is processed for a time period called a quantum. If during this quantum the job is completed, that job departs and service begins on the next; otherwise, the uncompleted job rejoins the system of queues to await further service. In some systems, priorities are assigned to customers. These priorities can be assigned externally [4]· they can be assigned to the customers as functions of their attained service time (the amount of service time so far obtained by a customer) [5]; or they can be assigned as functions of their waiting time [6], etc. Such priority queueing systems are called preemptive if the customer in the service facility is preempted whenever there exists another customer in the system who has higher priority.



GENERAL FEEDBACK QUEUEING MODEL

FIGURE 1-1

It is necessary to specify the arrival and the service processes before any analysis can be carried out. Let $A(t)$ denote the distribution function of the interarrival times with average time $1/\lambda$ seconds. If the interarrival times are exponentially distributed as

$$A(t) = P[\text{interarrival time} \leq t] = 1 - e^{-\lambda t} \qquad t \geq 0 \qquad (1.1)$$

then the arrival process is called Markovian (Poisson). Otherwise, it is called general. Also let us use $B(x)$ to denote the distribution function of the service times with mean request equal to $1/\mu$ seconds. If the service times are exponentially distributed as

$$B(x) = P[\text{service time} \leq x] = 1 - e^{-\mu x} \qquad x \geq 0 \tag{1.2}$$

then the service process is referred to as Markovian (exponential) to differentiate it from the general case.

Usually, two letters and a number are used to specify the arrival and service processes as well as the number of servers in the system. The first letter is used to specify the arrival process, the second letter for the service process, and the number is used to designate the number of servers in the system. The letter M is used for the Markovian process, and the letter G is used to represent a general process. All of the models to be analyzed in this dissertation are of either M/M/1 or M/G/1 type, namely, there is one server (CPU) in the system; the arrival process is Markovian (Poisson); and the service process is either exponential or general.

The utilization factor ρ, representing the percentage of time that the system is busy, is defined as the ratio of the average arrival rate and the average service rate.

$$\rho \equiv \frac{\lambda}{\mu} \tag{1.3}$$

ρ has to be smaller than 1 so that the average work load offered to the processor is less than its capacity to handle such a load.

Another interesting quantity in the system is the size of the quantum which is defined as the time interval allocated to a customer when he enters the service facility. In a real system, the quantum size has to be finite in order to get any work done, but the analysis tends to be difficult and the results tend to be in complex form under this assumption [1]. In 1967, the notion of allowing the quantum to shrink to zero was first studied [4] and is referred to as "processor-sharing." As the name implies, this zero-quantum limit provides a share or portion of the processing unit to many customers simultaneously. Under the assumption of processor-sharing, the difficulty in analysis disappears in large part and the results tend to be in simpler form. Of course, this assumption of infinitessimal quantum can never be reached in practice due to the consideration of overhead time; nevertheless, it usually can serve as a good approximation of the actual systems.

CHAPTER 2

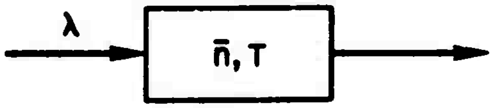
QUEUEING THEORY TOOLS AND SUMMARY OF ANALYTIC RESULTS FOR TIME-SHARED SYSTEMS

2.1 Queueing Theory Tools

Queues were first studied systematically by Erlang [7]. Others who have made key contributions to the mathematical theory of queues are Pollaczeck [8,9], Kolomogorov [10], Kendall [11,12], Lindley [13], and Takacs [14-17], to mention a few. Mathematical models of time-sharing systems are stochastic in nature and their analysis thus draws heavily on queueing theory results. In this section, we present some of the queueing theory results that will be used later in this dissertation for the analysis of our mathematical models.

2.1.1 Little's Result [18]

Let $\bar{n}$ denote the expected number of customers in a queueing system and T the expected time that they spend in the system. Assume that the average rate of arrival is λ. Refer to Figure 2-1.



A GENERAL QUEUEING SYSTEM

FIGURE 2-1

We assume that the box is "conservative" in the sense that customers are neither created nor destroyed nor on the average accumulated within that system, and so clearly, the average departure rate must be λ. Under the constraint that the stochastic process involved is erogodic [19], Little [18] proved that the following relationship is always true.

$$\bar{n} = \lambda T \qquad (2.1)$$

2.1.2 Memoryless Property of the Markovian Process [19]

As the name indicates, the past history of a random variable which is distributed exponentially in no affects its future. The following equation expresses this property for a random variable T.

$$P[T \leq t + t_0 | T > t_0] = 1 - e^{-\lambda t} = P[T \leq t] \qquad t \geq 0 \qquad (2.2)$$

The distribution of time until the next event (say, an arrival) occurs given that t_0 seconds have elapsed since the occurrence of the last event is identically equal to the distribution of time until the next event occurs measured from the time when the last event occurred. Thus the time a future event occurs is independent of how long it has been since the last event occurred. In other words, the Markovian process is memoryless.

2.1.3 Markovian Process (M/M/1) [19]

For an M/M/1 system with infinite queueing room, since both the arrival and the service process are Markovian, the all-important memoryless property holds, and the results of birth-death processes can be applied directly. For such systems, the equilibrium probability of having n customers in the system is given as

$$p_n = (1 - \rho)\rho^n \qquad n = 0,1,2,... \qquad (2.3)$$

where ρ is the utilization factor defined by Eq. (1.3), and $\rho < 1$.

The expected number of customers in the system $\bar{n}$ can be calculated as

$$\bar{n} = \sum_{n=0}^{\infty} np_n = \frac{\rho}{1 - \rho} \tag{2.5}$$

We may now apply Little's result in order to obtain the average time spent in the system as follows:

$$T = \frac{\bar{n}}{\lambda} = \frac{1/\mu}{1 - \rho} \tag{2.6}$$

2.1.4 The Imbedded Markov Chain (M/G/1) [12]

For general service time distributions the nice property of memorylessness no longer exists, and the results of birth-death processes can no longer be applied directly to the system. However, when the system is studied at discrete time points, the collection of state probabilities may constitute a Markov chain. Kendall [12] introduced the concept of an imbedded Markov chain so that a non-Markovian process can be studied by extracting a set of points (called regeneration points) at which the Markov property holds. For a M/G/1 system, the set of departure instants from service is an extremely convenient set of regeneration points. It is clear that if we specify the number of customers left behind by a departing customer, we can calculate the same quantity at some point in the future given only additional inputs to the system. From the analysis of the imbedded Markov chain, we get the following two important results:

A. Pollaczek-Khinchin Formula [19]

The average number of queueing customers (those customers waiting in the queue) left behind by a departing customer is given as

$$\bar{q} = \rho + \rho^2 \frac{(1 + C_b^2)}{2(1 - \rho)} \tag{2.6}$$

where C_b is the coefficient of variation defined as the ratio of the standard deviation σ_b of the service time distribution to its mean value.

$$C_b = \frac{\sigma_b}{1/\mu} = \mu\sigma_b \qquad (2.7)$$

If we apply Little's result to Eq. (2.6), we get the average time spent in the system as

$$T = \frac{\bar{q}}{\lambda} = \frac{1}{\mu} + \frac{\rho}{\mu} \cdot \frac{(1 + C_b^2)}{2(1 - \rho)} \qquad (2.8)$$

Equation (2.8) is easily interpreted. The average total time spent in the system is clearly the average time spent in the service plus the average time spent in the queue.

B. Distribution of Waiting Time [19]

$$Q(z) = B^*(\lambda - \lambda z) \frac{(1 - \rho)(1 - z)}{B^*(\lambda - \lambda z) - z} \qquad (2.9)$$

where $Q(z)$ is the z transform of the distribution function of the number of customers in the queue. Let p_n denote the stationary probability that there are n customers in the queue, then $Q(z)$ is defined as

$$Q(z) = \sum_0^\infty p_n z^n \qquad (2.10)$$

$B^*(s)$ is the Laplace transform of the service time density function $\frac{dB(x)}{dx}$ defined by $B^*(s) = \int_0^\infty e^{-sx} dB(x)$.

From Eq. (2.9) the Laplace transform $S^*(s)$ of the distribution of total time spent in the M/G/1 system can be obtained as

$$S^*(s) = B^*(s) \frac{s(1 - \rho)}{s - \lambda + \lambda B^*(s)} \qquad (2.11)$$

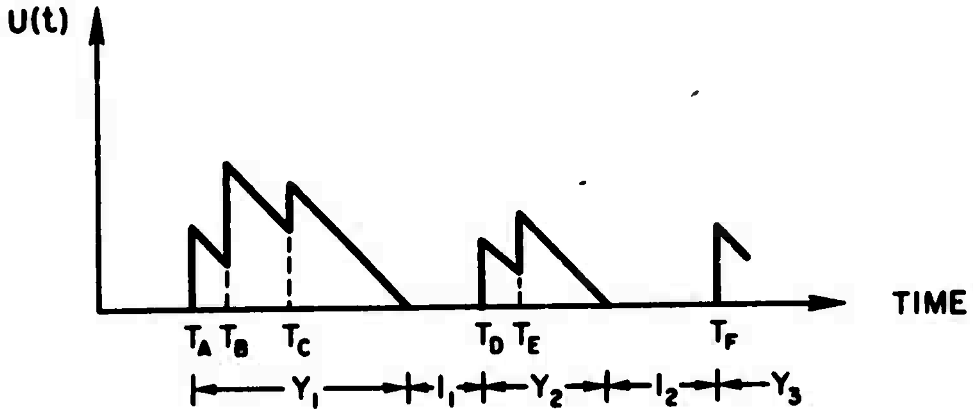
Since the waiting time in the queue for a customer is independent of his own service time, we easily get the Laplace transform of the waiting time as

$$W^*(s) = \frac{s(1 - \rho)}{s - \lambda + \lambda B^*(s)} \tag{2.12}$$

Differentiating Eqs. (2.11) and (2.12), various moments of the waiting time and the system time can be derived.

2.1.5 The Busy Period [19]

The queueing system can be viewed as passing through alternating cycles of busy periods and idle periods as depicted by Figure 2-2.



THE UNFINISHED WORK AND THE BUSY PERIOD

FIGURE 2-2

In this figure we plot

$U(t)$ = the unfinished work in the system at time t

= the time required to empty the system of all customers present at time t.

We assume that customers arrive at time epochs $T_A, T_B, T_C, \dots$. Each arrival to the system will add some unfinished work to the system (namely, his service time) as shown by Figure 2-2. U(t) is sometimes referred to as the virtual waiting time at time t. Behavior of this function is extremely important in understanding queueing systems when one views them from the point of view of the busy period. In Figure 2-2, $Y_1, Y_2, \ldots$ represents busy periods, and $I_1, I_2, \ldots$ represents idle periods.

For M/G/1 systems in general, since the arrival time distribution is memoryless, the moments of the idle period are the same as the moments of the interarrival time, namely,

$$F(r) = P[\text{idle period time} \leq r] = 1 - e^{-\lambda r} \qquad r \geq 0 \tag{2.11}$$

The analysis for the busy period distribution is much more complicated. The result is given by the following recursive equation [19].

$$P^*(s) = B^*[s + \lambda - \lambda P^*(s)] \tag{2.12}$$

where $P^*(s)$ is defined as the Laplace transform of the distribution function of the busy period. From Eq. (2.12) the first two moments of the length of the busy period can be obtained as

$$g_1 = \frac{1/\mu}{1 - \rho} \tag{2.13}$$

$$g_2 = \frac{\overline{t^2}}{(1 - \rho)^3} \tag{2.14}$$

where $\overline{t^2}$ is the second moment of the service time distribution. Comparing Eqs. (2.13) and (2.6) we find that the average length of a busy period for the system M/G/1 is equal to the average time a customer spends in an M/M/1 system.

2.2 Review of Some Analytic Results for Time-Shared Systems

In this section we wish to present some of the analytic results that were obtained in the past. They served as a point of departure for the research of this dissertation. The emphasis of this collection of results is on those for the processor-sharing models, although some of the relevant results for the finite-quantum systems are also presented. Typically, the quantity that is solved for in a time-sharing system is the distribution of the response time which is defined as the total time a customer spends in the system conditioned on that he requests and gets t seconds of service. Most of the time, however, we can only solve for the average response time defined as

$$T(t) = E[\text{response time for a customer conditioned on that he requests } t \text{ seconds of service}] \tag{2.15}$$

Another quantity, W(t), defined as the average amount of wasted (waiting) time spent in the system, is also often used as a performance measure for time-sharing systems, clearly

$$W(t) = T(t) - t \tag{2.16}$$

Swap time is assumed to be negligible for the following results; its effect on the response time can often be taken into consideration by reducing the average service rate of the service facility [20].

2.2.1 First-Come-First-Served (FCFS) System

This system is also known as batch processing. New comers always join the tail of the queue (there is only one queue in this system) and once a customer enters the service, he will be served until completion. We can regard this as a special case of time-sharing systems with infinite quantum size. The average response time for this case with general service time distribution is [19]

$$T(t) = \frac{\lambda \overline{t^2}}{2(1 - \rho)} + t \qquad (2.17)$$

and

$$W(t) = \frac{\lambda \overline{t^2}}{2(1 - \rho)} \qquad (2.18)$$

where $\overline{t^2}$ is the second moment of the service time distribution. A very important characteristic of $W(t)$ is that it is independent of t. For the system of exponentially distributed service, Eq. (2.18) becomes

$$W(t) = \frac{\rho/\mu}{1 - \rho} \qquad (2.19)$$

2.2.2 Last-Come-First-Served (LCFS) System

In this system a newly arrived customer captures the use of the server until he leaves completely served or until he is preempted by a newly arriving customer. At all times, the customer who has been in the system for the least amount of time will occupy the service. No more than one customer can be in the service at any time. The average response time for the LCFS system is given by [19]

$$T(t) = \frac{t}{1 - \rho} \qquad (2.20)$$

2.2.3 Shortest Job First Served (SJF) System

In this system the server selects the customer in the queue with the shortest required service time and serves it until completion. This algorithm requires the knowledge of the service time request in advance which is usually assumed to be unavailable in other algorithms. The average response time with general service distribution is given as

[21]

$$T(t) = \frac{\frac{1}{2}\lambda\overline{t^2}_{<t}}{(1 - \lambda\overline{t}_{<t})^2} \tag{2.21}$$

with $\overline{t^2}_{<t}$ and $\overline{t}_{<t}$ defined by Eq. (2.51).

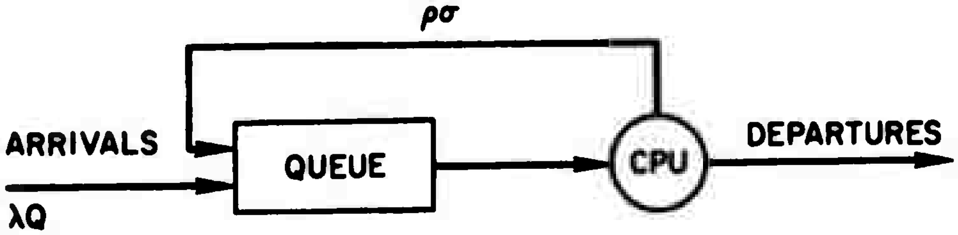
2.2.4 Round Robin (RR) System

A. Finite-Quantum

This discrete time model was first studied by Kleinrock [22]. The system works in the following way: arriving customers are queued in order of arrival. The server selects the customer at the head of the queue and services him for at most Q seconds, where Q is the quantum size. If after Q seconds of service the customer needs more, he is returned to the end of the same queue. The service time of a newly arriving customer is chosen independently from a geometric distribution such that for $\sigma < 1$,

$$S_n = (1 - \sigma)\sigma^{n-1} \qquad n = 1,2,3,... \tag{2.22}$$

where S_n is the probability that a customer's service request is exactly nQ seconds. See Figure 2-3.



THE DISCRETE TIME ROUND ROBIN MODEL

FIGURE 2-3

At the end of each time interval, a new customer arrives to the system with probability Q; thus, the average arrival rate is λ. There are two types of systems dependent upon the order in thich a new arrival and the ejection of the customer in service can take place at the end of each quantum. The average response time for a customer needing nQ seconds in the early-arrival system is given by [22]

$$T(nQ) = \frac{nQ}{1 - \rho} - \rho Q - \frac{\lambda Q^2 \rho}{1 - \rho} [1 + \frac{(1 - \sigma\alpha)(1 - \alpha^{n-1})}{(1 - \alpha)^2 (1 - \rho)}] \tag{2.23}$$

where

$$\alpha = \sigma + \lambda Q \tag{2.24}$$

$$\rho = \frac{\lambda Q}{1 - \sigma} \tag{2.25}$$

Similarly, the average response time for a customer requesting nQ seconds of service in the late-arrival system is given by [22]

$$T(nQ) = \frac{nQ}{1 - \rho} - \frac{\lambda Q^2}{1 - \rho} [1 + \frac{(1 - \sigma\alpha)(1 - \alpha^{n-1})}{(1 - \alpha)^2 (1 - \rho)}] \tag{2.26}$$

B. Processor-Sharing

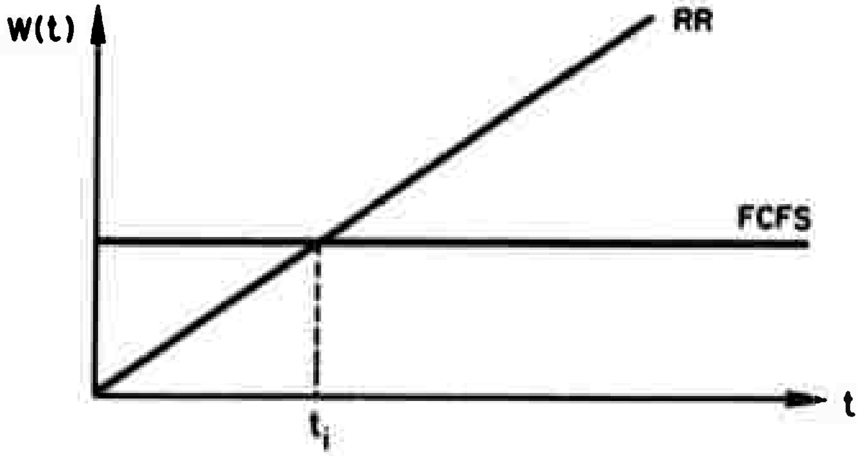
As the quantum size shrinks, customers get served at a faster rate but with less service each time. For the limit case of zero quantum, a customer is required to make an infinite number of cycles, each infinitely quickly and each time receiving infinitessimal service, until he finally accumulates enough service time (to be equal to his request), at which time he leaves. The average time for such a processor-sharing RR system is given as

$$T(t) = \frac{t}{1 - \rho} \tag{2.27}$$

and

$$W(t) = \frac{\rho t}{1 - \rho} \tag{2.28}$$

Kleinrock [4] obtained this for the case of exponential service distribution. Sakata [23,24] proved that it is also true for general service distribution. Coffman, Muntz, and Trotter [25] solved for the Laplace transform of the waiting time distribution for the system M/M/1.



WAITING TIME FUNCTION FOR THE FCFS AND RR SYSTEMS

FIGURE 2-4

In Figure 2-4 we plot W(t) against t for the FCFS and the RR systems. Both of them are straight lines. Let us assume that these two lines intersect each other at the point $t = t_i$. This point t_i is of great interest to us because for a customer requesting less than t_i seconds of service, then he has to wait more than the average as represented by FCFS when he is in a RR system. t_i can easily be calculated as

$$\frac{\rho t_i}{1 - \rho} = \frac{\lambda \overline{t^2}}{2(1 - \rho)} \tag{2.29}$$

$$t_i = \frac{\mu \overline{t^2}}{2} \tag{2.30}$$

For exponential service distribution, $\overline{t^2} = \frac{2}{\mu^2}$, thus,

$$t_i = \frac{1}{\mu} \tag{2.31}$$

For M/M/1 systems, any customer who needs more than the average service request (1/μ seconds) will be better off as measured by his average response time when he is in the FCFS system than in the RR system.

2.2.5 Round Robin with Priorities System

In this model we assume that an external priority assignment is made to the arriving customers. We assume that there are P priority groups with Poisson arrivals, each at an average rate λ_p per second and an exponentially distributed service requirement with mean request of $1/\mu p$ seconds for the p^{th} group. A processor-sharing model is assumed. A positive number g_p which denotes the relative fraction of the processing time that is reserved for the customers from the p^{th} priority group is associated with the p^{th} priority group, with larger values of g_p being given to those higher priority groups. The average response time $T_p(t)$ for a customer from the p^{th} priority group is given by [4]

$$T_p(t) = \frac{t}{1 - \rho} [1 + \sum_{i=1}^{P} (\frac{g_i}{g_p} - 1)\rho_i] \qquad p = 1,2,\ldots,P \tag{2.32}$$

where

$$\rho_i = \lambda_i/\mu_i \tag{2.33}$$

$$\rho = \sum_{i=1}^{p} \rho_i \tag{2.34}$$

2.2.6 Selfish Round Robin (SRR) System

This system was introduced by Kleinrock [26] for the processor-sharing model only. He solved for the mean response time for the case of exponential service time distribution. The algorithm works in the following way: For each customer in the system, a time-varying value of priority is assigned. This priority value begins at zero upon his entry to the system. It increases at a positive rate α as long as he is not served; whenever he is in the service facility, his priority value increases at a positive rate β where $\alpha \geq \beta \geq 0$. All the customers in the service facility share it equally among them as in a processor-sharing RR system. Note that a queueing customer gains priority at a rate greater than those in the service. Eventually he will catch up with those in service and then join and remain with that group. Since the customers in service are attempting to run away with the processor and prevent waiting customers from joining them (there attempt is futile, though), this system is called selfish round robin.

The average response time for the SRR system is [26]

$$T(t) = \frac{1/\mu}{1 - \rho} + \frac{(t - 1/\mu)}{1 - \rho(1 - \beta/\alpha)} \tag{2.35}$$

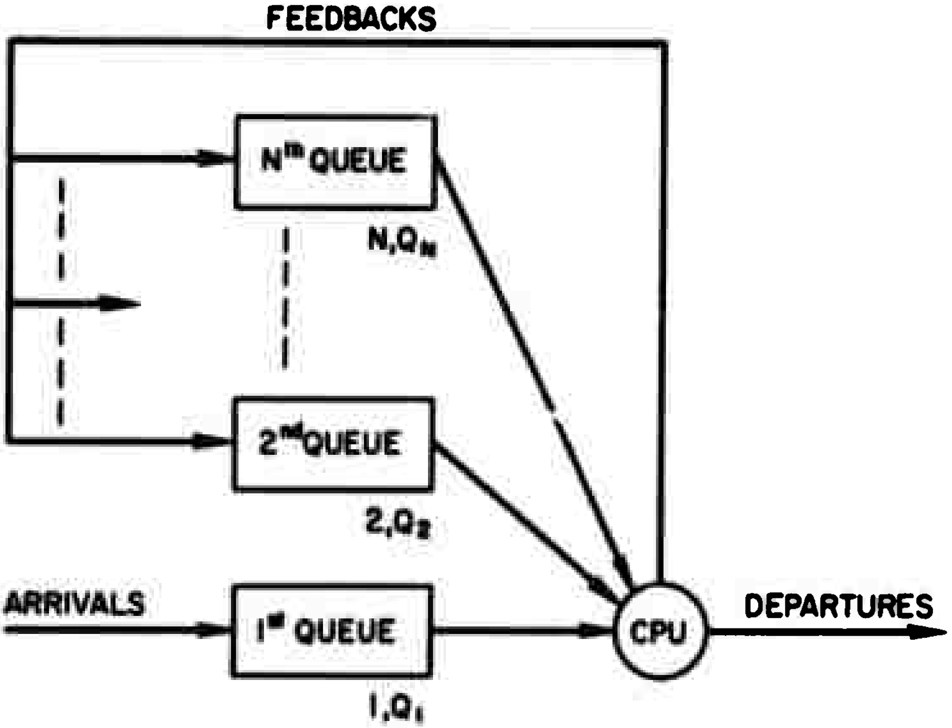
and also the average waiting time is given by

$$W(t) = \frac{\rho/\mu}{1 - \rho} + \frac{(t - 1/\mu)\rho(1 - \beta/\alpha)}{1 - \rho(1 - \beta/\alpha)} \qquad (2.36)$$

The ratio β/α provides one degree of freedom which can be adjusted over a continuum of system behaviors ranging from the FCFS $(\beta/\alpha = 1)$ to the RR system $(\beta/\alpha = 0)$. Another important property of the SRR systems is that a job with average service requirement will receive the same response in all of these SRR systems. Please refer to Chapter 3 for more details.

2.2.7 Foreground-Background (FB) System

A. Finite Quantum



DISCRETE TIME FOREGROUND-BACKGROUND MODEL

FIGURE 2-5

We assume that there are N levels of queues in a FB system as depicted by Figure 2-5. Arriving customers enter the level 1 queue to await allocation of their first quantum, Q_1 seconds. If more processing time is needed, they enter the end of the level 2 queue to await the second allocation, this time of Q_2 seconds. This process continues until either the customer leaves the system completely served or he enters the N^{th} queue. Members of the N^{th} queue are served as in a RR system. The server services a customer from the I^{th} queue only if all lower level queues $(I - 1, I - 2, \ldots, 1)$ are empty. If a new customer enters level 1 during execution of a customer from a higher queue, the current customer is not preempted until the allocated quantum is expired. The average response time with $Q_1 = Q_2 = \ldots = Q_N$ has been derived by Coffman and Kleinrock [5] as

$$T(t) = \frac{\rho/\mu}{(1 - \rho)[1 - \rho(1 - e^{-\mu(N-1)Q}]}$$

$$+ \frac{\rho[1 - e^{-\mu(N-1)Q}]}{1 - \rho[1 - e^{-\mu(N-1)Q}]}(K - 1)Q + t \qquad K \geq N \qquad (2.37)$$

$$T(t) = \frac{(\lambda/2)[E_K(\tau^2) + \gamma_K E_1(\tau^2)]}{[1 - \rho(1 - e^{-\mu KQ})][1 - \rho(1 - e^{-\mu(K-1)Q}]}$$

$$+ \frac{\rho[1 - e^{-\mu(K-1)Q}]}{1 - \rho[1 - e^{-\mu(K-1)Q}]}(K - 1)Q + t \qquad 1 \leq K < N \qquad (2.38)$$

where

$$(K - 1)Q < t \leq KQ \qquad (2.39)$$

$$\gamma_K = \frac{e^{-\mu KQ}}{1 - e^{-\mu Q}} \tag{2.40}$$

$$E_K(\tau^2) = \int_0^\infty \tau^2 dF_k(\tau) \tag{2.41}$$

$$F_K(\tau) = \begin{cases} 1 - e^{-\mu\tau} & 0 \leq \tau \leq KQ \\ 1 & \tau > KQ \end{cases} \tag{2.42}$$

Schrage [27] has provided analysis of this model with general service distribution and $N = \infty$. In particular, he solved for the Laplace transform of the response time distribution under the assumption of arbitrary quantum size for each level. Schemer [28] also contributed to the infinite-level FB model by obtaining the average response time as

$$T(t) = t + \sum_{i=1}^{K} \tau_i \tag{2.43}$$

where K is defined by the following inequality

$$\sum_{i=1}^{K-1} Q_i < t \leq \sum_{i=1}^{K} Q_i \tag{2.44}$$

and τ_i is given as

$$\begin{aligned} \tau_i = \bar{Q}_1(\tau_{i-1} + Q_{i-1})\lambda + \bar{Q}_2(\tau_{i-2} + Q_{i-2})\lambda e^{-\mu Q} + \ldots \\ + \bar{Q}_j(\tau_{i-j} + Q_{i-j})\lambda e^{-\mu tj-1} + \ldots + \bar{Q}_A \lambda e^{-\mu(i-1)\bar{Q}_B} \end{aligned} \tag{2.45}$$

with

$$\bar{Q}_i = \frac{1}{\mu}[1 - e^{-\mu Q_i}] \tag{2.46}$$

$$\bar{Q}_A = \frac{1}{\mu}[1 + \sum_{i=1}^{\infty} i \int_{t_i}^{t_{i+1}} \mu e^{-\mu t} dt]^{-1} \tag{2.47}$$

$$\bar{Q}_B = Q_1(1 - e^{-\mu Q_1}) + \sum_{i=2}^{\infty} Q_i (e^{-\mu t_{i-1}} - e^{\mu t_i}) \tag{2.48}$$

$$t_i = \sum_{j=1}^{i} Q_j \tag{2.49}$$

B. Processor-Sharing

Coffman and Kleinrock [5] and Schrage [27] independently derived the average response time for this case. With general service time distribution, $T(t)$, is given by

$$T(t) = \frac{W_{<t} + t}{1 - \rho_{<t}} \tag{2.50}$$

where

$$\overline{t^n_{<t}} = \int_0^t x^n dB(x) + x^n \int_t^{\infty} dB(x) \tag{2.51}$$

$$\rho_{<t} = \lambda \bar{t}_{<t} \tag{2.52}$$

$$W_{<t} = \frac{\lambda \overline{t^2_{<t}}}{2(1 - \rho_{<t})} \tag{2.53}$$

Schrage [27] also obtained the Laplace transform of the response time function for M/G/1 systems. Refer to Chapter 3 for more details.

2.2.8 Multilevel Processor-Sharing Systems

Multilevel (ML) queueing models were first analyzed by Kleinrock and Muntz [29].* They can be considered as a generalization and consolidation of the FCFS, the RR, and the FB systems. In particular, a set of attained service times $\{a_i\}$ is defined such that

$$0 = a_0 < a_1 < a_2 < \ldots < a_N < a_{N+1} = \infty \qquad (2.54)$$

The discipline for a job when it has attained service, τ, in the interval

$$a_{i-1} \leq \tau < a_i \qquad i = 1,2,3, \ldots, N + 1 \qquad (2.55)$$

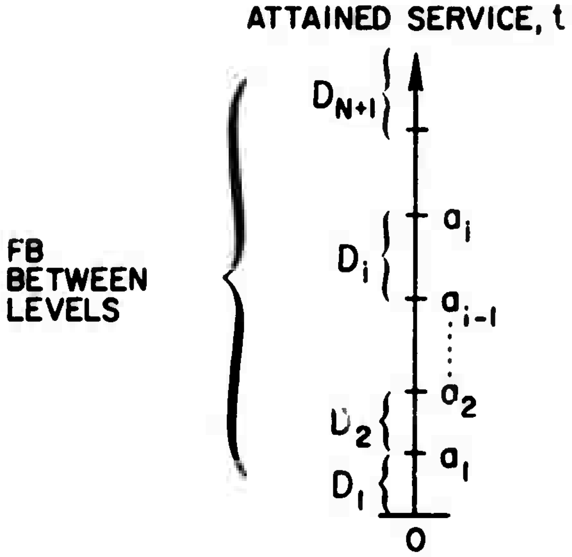
will be denoted as D_i. Where D_i is considered for any given level to be either FCFS, FB, or RR. Moreover, between intervals the jobs are treated as a set of FB disciplines. The behavior of the average conditional response time in any particular level is independent of the discipline in all other levels. See Figure 2-6.

An expression for $T(t)$, the mean response time for jobs with service time t such that $a_{i-1} < t \leq a_i$, i.e., jobs which reach the i^{th} level queue and there leave the system, has been obtained by Kleinrock and Muntz as [29]

a. i^{th} level discipline is FB

$$T(t) = \frac{t}{1 - \rho_{<t}} + \frac{\lambda \overline{t^2}_{<t}}{2(1 - \rho_{<t})^2} \qquad (2.56)$$

*Two other not yet published papers by these authors are "Processor-Sharing Queueing Models of Mixed Scheduling Disciplines" and "The Processor-Sharing Queueing Model for Time-Shared Systems with Bulk Arrivals" (with E. Rodemich).



INTERVALS OF ATTAINED SERVICE WITH DISCIPLINES, D_i

FIGURE 2-6

b. i^{th} level discipline is FCFS

$$T(t) = \frac{W_{<a_i} + t}{1 - \rho_{<a_{i-1}}} \tag{2.57}$$

where

$$W_{<a_i} = \frac{\lambda \overline{t^2}_{<a_i}}{2(1 - \rho_{<a_i})}$$

c. i^{th} level discipline is RR

In this case, the results are limited in the i^{th} interval to service time distribution in which

$$B(x) = 1 - p(x)e^{-\beta x} \qquad a_{i-1} \leq x < a_i \tag{2.58}$$

$$p(x) = p_0 + p_1 x + \ldots + p_n x^n \tag{2.59}$$

The service time distribution $F(x)$ for this i^{th} level is then

$$F(x) = \begin{cases} \dfrac{B(a_{i-1} + x) - B(a_{i-1})}{1 - B(a_{i-1})} = 1 - q(x)e^{-\beta x} & 0 \leq x < a_i - a_{i-1} \\ 1 & x \geq a_i - a_{i-1} \end{cases} \tag{2.60}$$

where

$$q(x) = \frac{e^{- a_{i-1}} p(a_{i-1} + x)}{1 - B(a_{i-1})} = q_0 + q_1 x + \ldots + q_n x^n \tag{2.61}$$

Except for the first level the average response time is given as

$$T(a_{i-1} + \tau) = \frac{1}{1 - \rho_{<a_{i-1}}} \{W_{<a_{i-1}} + a_{i-1} + \alpha_2(\tau)\} \tag{2.62}$$

where

$$\alpha_2(\tau) = \frac{\tau}{k - \lambda\bar{a}\,\frac{1}{\mu_i}} + \frac{b(\mu^2 - \gamma_1^2)[(\gamma_1 + \mu - \lambda\bar{a})(1 - e^{-\gamma_1\tau}) - \lambda\bar{a}\,\bar{e}^{-(\mu+\gamma_1)x_1}(e^{\gamma_1\tau} - 1)]}{2\lambda\bar{a}\gamma_1^2[\gamma_1 + \mu - \lambda\bar{a}(1 - e^{-(\mu+\gamma_1)x_1})]} \tag{2.63}$$

$$W_{<a_{i-1}} = \frac{\lambda(1 - e^{-\mu a_{i-1}} - \mu a_{i-1}e^{-\mu a_{i-1}})}{\mu^2[1 - \frac{\lambda}{\mu}(1 - e^{-\mu a_{i-1}})]} \qquad (2.64$$

with

$$\bar{a} = \frac{1 - B(a_{i-1})}{1 - \lambda\bar{t}_{<a_{i-1}}} \tag{2.65}$$

$$\mu_i = \frac{\mu}{(1 - e^{-\mu x_1})} \tag{2.66}$$

$$\gamma_1 = \sqrt{\mu^2 - 2\mu\lambda\bar{a} + (\lambda\bar{a})^2(1 - e^{-2\mu x_1})} \tag{2.67}$$

and

$$x_1 = a_i - a_{i-1}$$

For the first level, $T(t)$ is given as

$$T(t) = \frac{t}{1 - \rho_{<a_1}} \qquad 0 \leq t < a_1 \tag{2.68}$$

In Figure 2-7 we show the behavior of each of the three disciplines for the system $N = 1$ with exponential service distribution. We also assume that $\mu = 1$, $\lambda = 0.75$, and $a_1 = 2$.

2.2.9 Attained Service [30]

A. Finite-Quantum

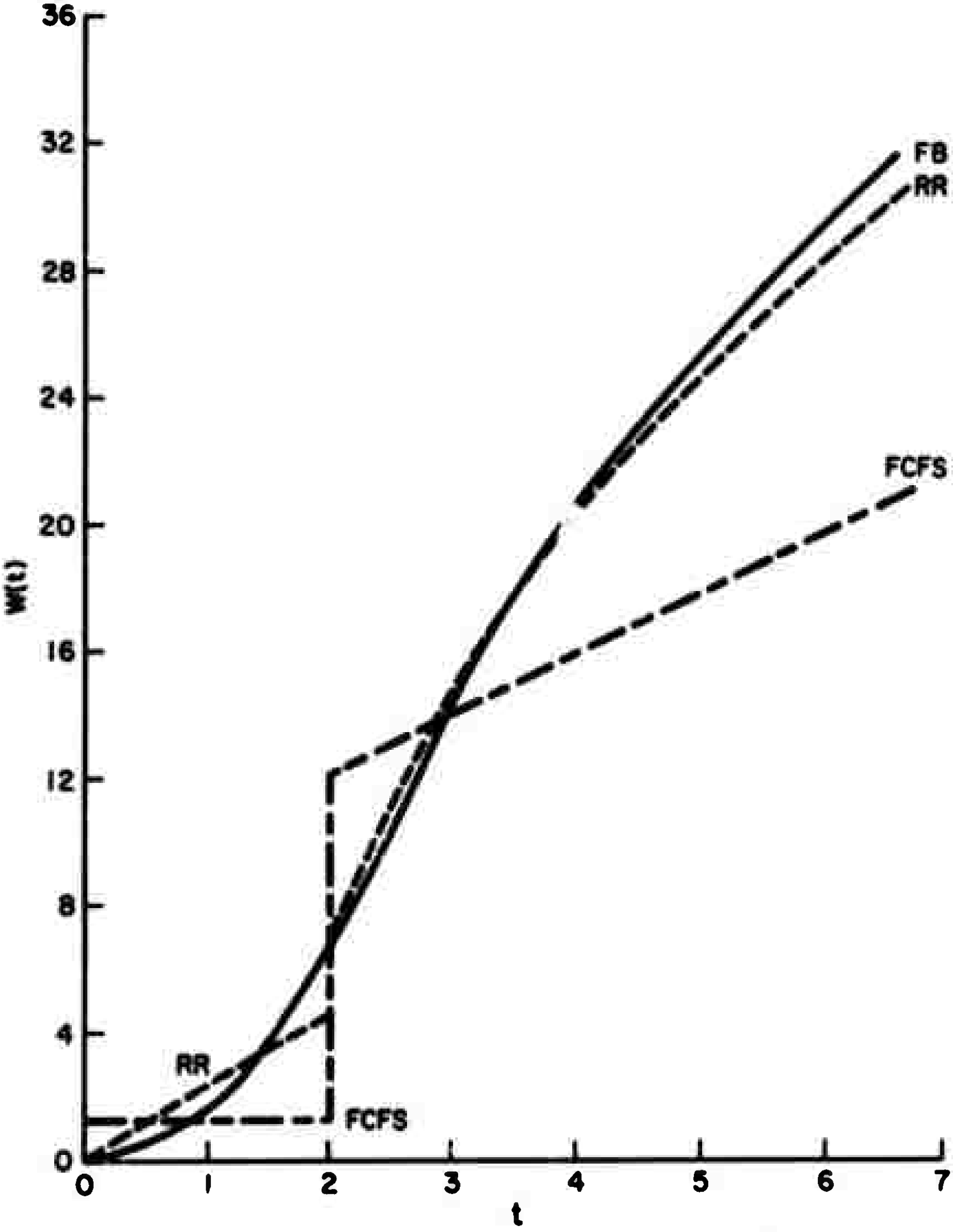
The attained service for an incompletely serviced customer is defined as the number of seconds that he has so far spent in the service facility. Assume that there are p priority groups in the system and let

$N_p(t)$ = density of the number of customers in the system from priority group p who have so far received exactly t seconds of service

λ_p = average arrival rate of the p^{th} priority group

$B_p(t)$ = service distribution of the customers from the p^{th} priority group

$T_p(t_n)$ = average response time for customer from p^{th} group conditioned on that he has attained t_n seconds of service



AVERAGE WAITING TIMES FOR N=1, M/M/1, $\bar{t}$= 1, λ= 0.75

FIGURE 2-7

Kleinrock then proved that the following equation is true

$$N_p(t_n) = \lambda_p[1 - B_p(t_n)][T_p(t_{n+1}) - T_p(t_n)] \qquad p = 1,2,3, \ldots P \quad (2.69)$$

B. Processor-Sharing

If we define

$n_p(t)$ = average density of customers from the p^{th} priority group still in the system who have so far received t seconds of service

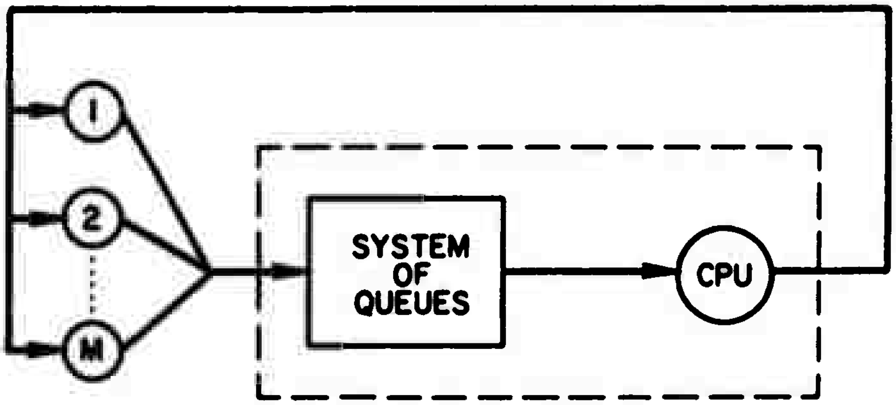
then Eq. (2.55) can be modified for the process-sharing models as [30]

$$n_p(t) = \lambda_p[1 - B_p(t)] \frac{dT_p(t)}{dt} \qquad p = 1,2,3, \ldots, P \qquad (2.70)$$

2.2.10 Round Robin with Finite Input Population

In the real world, there is no such thing as an infinite population on which most of the mathematical models of time-sharing systems are based. Nevertheless, if the dependence of the arrival process upon the number of customers in the system is negligible, the assumption of infinite population usually serves as a good approximation to the real system. In some of the systems, however, the arrival process does depend on the number of customers in the system in a perceptible way. The analysis of such models with finite input population then becomes a necessity.

Typically, a time-varying model with finite number of inputs is modelled as by Figure 2-8. Here we have M users which make demands on the time-shared system. The dashed lines in Figure 2-8 surround a feedback queueing model similar to the one used for infinite population models. When a user (a console) makes a request for service to the computer, he enters the dashed box and gets served according to the



TIME-SHARING MODEL WITH M INPUT CONSOLES

FIGURE 2-8

scheduling algorithm in the system. When his request is completely served, he leaves the dashed box and starts to generate a new request to the CPU. The time spent by the user in generating this new request after the completion of the previous request is referred to as the "think time." Thus, alternating periods of thinking and processing take place.

Scherr [31] considered the case for which he assumes exponentially distributed service time and think time, namely

$$P[\text{think time} \leq t] = 1 - e^{-\alpha t} \qquad t \geq 0 \tag{2.71}$$

$$\text{average think time} = 1/\alpha \tag{2.72}$$

$$P[\text{service time} \leq x] = 1 - e^{-\mu x} \qquad x \geq 0 \tag{2.73}$$

$$\text{Average service request} = 1/\mu \tag{2.74}$$

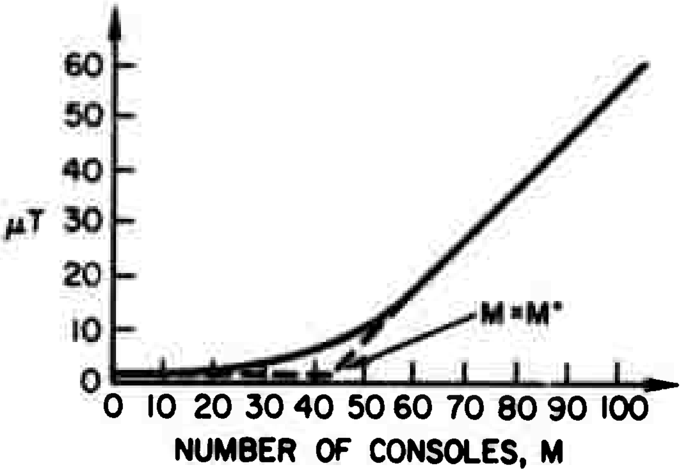
He solved for the average response time in the system without conditioning that result on the service time required. His result is

$$T = \frac{M/\mu}{1 - p_0} - \frac{1}{\alpha} \qquad (2.75)$$

where p_0 is the probability that no customer is in the queue or in the service facility and is given by

$$p_0 = \left[\sum_{m=0}^{M} \frac{M!}{(M-m)!} \left(\frac{\alpha}{\mu}\right)^m \right]^{-1} \qquad (2.76)$$

In Figure 2-9 we plot the normalized waiting time μT against the number of input population M. The point M^* on Figure 2-9 is defined by Kleinrock [32] as the saturation number because whenever the



FINITE POPULATION PERFORMANCE AND SATURATION

FIGURE 2-9

total number of consoles exceeds this number, every additional customer will severely interfere with the existing customers as far as the average response time is concerned. M^* can be easily calculated as

$$M^* = \frac{1/\mu + 1/\alpha}{1/\mu} = \frac{\mu + \alpha}{\alpha} \qquad (2.77)$$

Greenberg [33], Adiri and Avi-Itzhak [34], and Krishnamoorthi and Wood [35] also considered systems with finite input population.

2.2.11 A Conservation Law

Kleinrock [36] showed that a conservation law holds for any queueing system with priorities which satisfies the following restrictions:

a. there is a single server in the system

b. arrival process is Poisson, service process is arbitrary with arrival and service processes independent of each other

c. work can not be destroyed nor can it be created within the system

d. preemption allowed only if service process is exponential and then preemption must cause no losses.

The conservation law is that the weighted average of waiting times is a constant regardless of the scheduling discipline, namely,

$$\sum_{p=1}^{P} \rho_p W_p = \frac{\rho W_0}{1 - \rho} \qquad (2.78)$$

where P is the number of priority groups and

λ_p = average arrival rate for p^{th} priority group

$1/\mu p$ = average service request for a customer from p^{th} priority group

W_p = average waiting time for customers from p^{th} priority group

W_0 = average amount of work left in the service facility found by an arriving customer (independent of scheduling algorithm)

2.3 Summary of Results in this Dissertation

There are two major topics in this dissertation. The first theme is on the modeling and analysis of new models of time-shared computer systems; the emphasis is on models with some degrees of freedom which the system designer can use to adjust the system performance over a continuum of system behaviors. In order to provide those degrees of freedom to the system designer, some parameters have to be injected into the system. Different algorithms with this property are discussed in Chapters 3 and 4. The second major topic is the finding of some fundamental properties which apply to the average number of customers in the system and the average response time functions for a large class time-shared computer systems. Chapters 5 and 6 are devoted to the discussion of these topics.

Chapter 3 is devoted to the study of the family of selfish scheduling systems in general, with the selfish round robin (SRR) and the selfish foreground-background (SFB) as two illustrative examples. In the selfish system, customers are divided into two sets: those in the queue box waiting to be served, and those in the service box sharing the service facility in some fashion determined by the scheduling algorithm being used. When a customer is waiting in the queue box, his priority (a number) increases at a positive rate α; when he is in the service box, his priority increases at a positive rate β. We consider the case $\alpha \geq \beta \geq 0$. If the scheduling algorithm in the service box is RR (namely, everyone in the service box shares the facility equally among themselves), the system is called selfish round robin (SRR). If only those customers in the service box with the least amount of attained service are sharing the server, then the system becomes selfish foreground-background (SFB). The ratio β/α provides to the system designer a degree of freedom to control the system performance. For the SRR system, the average response time is solved for general service distribution, and the Laplace transform of the waiting time distribution is obtained for the M/M/1 system. The Laplace transform of the waiting time distribution for the SFB system is solved for M/G/1 in general.

In Chapter 4 we discuss a family of algorithms whose performance ranges between that of the RR system and the FB system. Similar to a processor-shared RR system, all customers in the system share the service facility simultaneously; but unlike the RR system, the customers do not share the facility equally among themselves, rather their share of the processor varies according to their amount of attained service time. A weighting function $g(t) = ge^{-gt}$ is given to define the scheduling algorithm such that a customer's rate of attaining service, given that he has attained t seconds of it, is directly proportional to $g(t)$. The more service a customer has accumulated, the slower he gets served in the service facility. A customer always gets some service even though he has spent a long time in the service facility. Thus this new scheduling algorithm shows more discrimination against long jobs than the RR system, but less discrimination than the FB system. The parameter g provides one degree of freedom to the system designer. With $g = 0$ the system becomes RR; with $g = \infty$ it becomes FB. The average response time for this family of systems is obtained in Chapter 4.

In Chapter 5, conditioned on the presence of a "tagged" customer, we find a simple relationship between the time-dependent average number of customers in the system and the average response time function for a large class of systems with Poisson arrival and exponential service processes. The result shows that the behavior of one customer can strongly influence the total number of customers in the system for all algorithms except RR, in which case the average number of customers in the system is a constant.

From the results obtained in Chapters 3 and 4, as well as from some published papers [1], we see that by slightly modifying the scheduler of a time-shared system, a different model can easily be constructed and corresponding analytical results can be obtained. This process can go on and on with no end in sight. Clearly, one is tempted to seek some order in these results. For example, do there exist any invariants in behavior? Can we bound the possible range of performance regardless of structures?, etc. In Chapter 6 we try to answer some of these important questions. Fortunately, we are able to state a monotonicity property, a conservation law, and tight upper and lower bounds on the system performance as measured by average response time. Examples of the tight bounds are given for the exponential, the hyperexponential, the 2-stage Erlangian, and the uniform service time distributions.

CHAPTER 3

THE FAMILY OF SELFISH SCHEDULING ALGORITHMS (SSA)

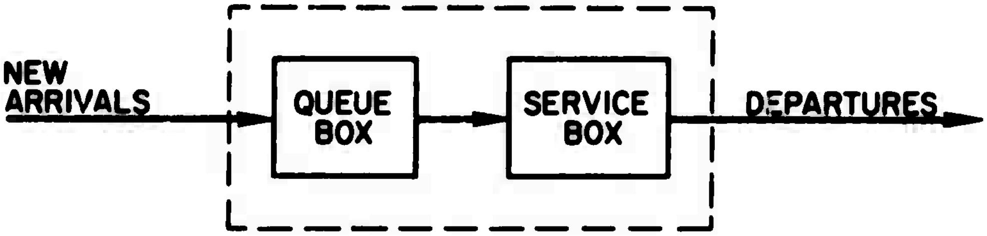
3.1 The Mathematical Model

The concept of selfish scheduling algorithms was first introduced by Kleinrock [26]. He solved the average response time for the selfish round robin (SRR) system. His work is extended and generalized in this chapter by obtaining the Laplace transform of the waiting time distribution for M/G/1 systems.

The principle behind this model is that all customers in the system are divided into two groups: those in a "queue box" waiting for service; and those in a "service box" sharing the service facility in a way as specified by the specific scheduling algorithm being used in the system. A newcomer always enters the queue box where his priority (a numerical value) increases from zero at a positive rate α; similarly, when he is in the service box (he may be sharing the service facility or he may be waiting for his turn depending upon the scheduler), his priority increases at a positive rate β. All customers possess the same parameters α and β. We are interested in the region $\alpha \geq \beta \geq 0$. Typically, a customer enters the queue box as soon as he arrives to the system, and starts to build his priority with rate α while he waits in the queue box. Since $\alpha \geq \beta \geq 0$, sooner or later he will catch up with those customers in the service box and join them to share the service facility. There is no feedback from the service box to the queue box, although there may be feedback within the service box. Kleinrock [26]

Preceding page blank

defined α as the queueing slope and β as the serving slope. If an RR scheduling algorithm is being used in the service, then everyone in the service box shares the service facility on an equal basis and the system is referred to as selfish round robin (SRR). If the server only serves those in the service box with the least amount of attained service as in an FB system, then the whole system becomes selfish foreground-background (SFB). Figure 3-1 shows a decomposition of the selfish system.



DECOMPOSITION OF THE SSA SYSTEM

FIGURE 3-1

Let us define the following quantities:

E_t : The event that customer needs t seconds of service

$s(t)$ = a random variable representing the total time a customer spends in the system conditioned on E_t.

$w(t)$ = a random variable representing the total time a customer wastes while waiting in the system conditioned on E_t.

$q(t)$ = a random variable representing the time a customer spends in the queue box conditioned on E_t.

$y(t)$ = a random variable representing the time a customer spends in the service box conditioned on E_t.

$v(t)$ = a random variable representing the time a customer wastes (the waiting time) in the service box conditioned on E_t.

$B(t)$ = $P[\text{service time} \leq t]$

$B^*(s)$ = The Laplace transform of the service time distribution $dB(x)$.

$$= \int_0^\infty e^{-sx} dB(x) \qquad (3.1)$$

and

$1/\mu$ = average service request.

λ = average arrival rate.

$\rho = \lambda/\mu$ = utilization factor of the system.

$S^*(t,s)$ = The Laplace transform of $s(t,x)$, the equilibrium density function of $s(t)$.

$$= \int_0^\infty e^{-sx} s(t,x)\, dx \qquad (3.2)$$

$W^*(t,s)$ = The Laplace transform of $w(t,x)$, the equilibrium tensity function of $w(t)$.

$$= \int_0^\infty e^{-sx} w(t,x)\, dx \qquad (3.3)$$

$Q^*(t,s)$ = The Laplace transform of $q(t,x)$, the equilibrium density function of $q(t)$.

$$= \int_0^\infty e^{-sx} q(t,x)\, dx \qquad (3.4)$$

$Y^*(t,s)$ = The Laplace transform of $y(t,x)$, the equilibrium density function of $y(t)$.

$$= \int_0^\infty e^{-sx} y(t,x)\, dx \qquad (3.5)$$

$V^*(t,s)$ = The Laplace transform of $v(t,x)$, the equilibrium density function of $v(t)$.

$$= \int_0^\infty e^{-sx} v(t,x)\,dx \tag{3.6}$$

Clearly we have that

$$s(t) = q(t) + y(t) \tag{3.7}$$

$$w(t) = q(t) + v(t) = s(t)-t \tag{3.8}$$

Let us also define

$$\begin{aligned} T(t) &= E\ [\text{response time}|E_t] \\ &= E\ [s(t)] \\ &= -\lim_{s\to 0} \frac{\partial S^*(t,s)}{\partial s} \end{aligned} \tag{3.9}$$

$$\begin{aligned} W(t) &= E\ [\text{wasted time in system}|E_t] \\ &= E\ [W(t)] \\ &= -\lim_{s\to 0} \frac{\partial W^*(t,s)}{\partial s} = T(t)-t \end{aligned} \tag{3.10}$$

$W_2(t)$ = second moment of the equilibrium waiting time distribution given E_t

$$= \lim_{s\to 0} \frac{\partial^2 W^*(t,s)}{\partial s^2} \tag{3.11}$$

$\sigma^2(t)$ = variance of the equilibrium waiting time distribution

$$= W_2(t) - W^2(t) \tag{3.12}$$

and

$$\begin{aligned} W_q(t) &= E\ (\text{waiting time in the queue box}|E_t) \\ &= E\ (q(t)) \\ &= -\lim_{s\to 0} \frac{\partial Q^*(t,s)}{\partial s} \end{aligned} \tag{3.13}$$

$$\begin{aligned} V(t) &= E\ [\text{waiting time in the service box}|E_t] \\ &= E\ [v(t)] \\ &= -\lim_{s\to 0} \frac{\partial V^*(t,s)}{\partial s} \end{aligned} \tag{3.14}$$

Since there is no feedback from the service box to the queue box, and all the service is done in the service box, the waiting time a customer spends in the queue box must be independent of his service request. Thus,

$$q(t) = q \tag{3.15}$$

$$Q^*(t,s) = Q^*(s) \tag{3.16}$$

$$W_q(t) = W_q \tag{3.17}$$

3.2 The Analysis of the SSA Systems

By solving the SSA system, we mean to find the equilibrium waiting time distribution of the system. Since we are unable to do that directly, we first obtain the Laplace transform of this distribution and then obtain the various moments by differentiation.

Before going into details of the analysis of the SSA system, we present the following well-known theorem for FCFS system; it will be used later.

Theorem 3.1 [19]

The Laplace transform of the equilibrium density function of the waiting time for the FCFS system is given as

$$W^*(t,s) = \frac{s(1 - \rho)}{\lambda B^*(s) - \lambda + s} \tag{3.18}$$

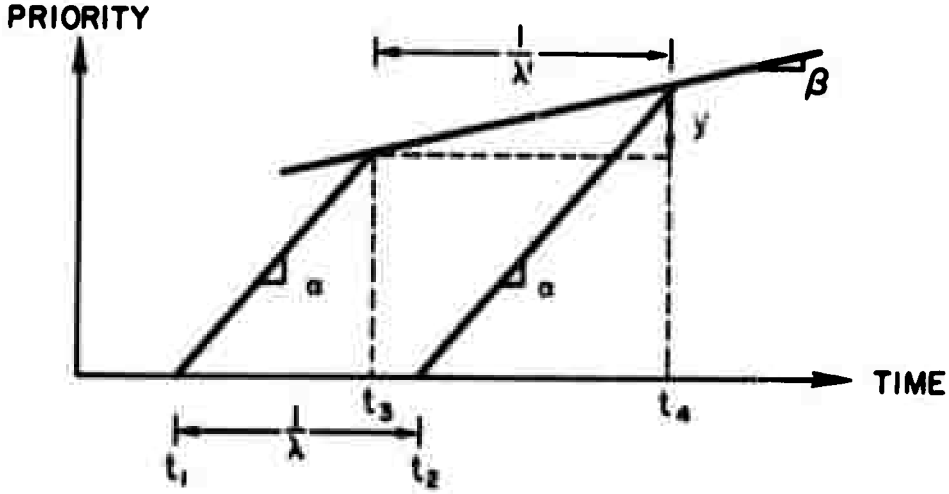
which is independent of t. Here $\rho = \lambda/\mu$

When an FCFS scheduling algorithm is used, there can be, at most, one customer being served in the service box at any time (no processor-sharing takes place in this case). The time a customer spends in the queue box is independent of the time he spends in the service box (t seconds).

Let us look at the service box of the SSA system as depicted by Figure 3-1. In order to solve $V^*(t,s)$, let us follow a customer, which we shall refer to as the "tagged" customer, through the system given that this customer requires t seconds of service. The arrival rate of customers to the service box conditioned on the presence of a tagged customer in that box is no longer λ, but rather some new average arrival rate λ', although the arrival process is still Poisson [26]. Thus (in conjunction with the fact that work can be done to a customer only when he is in the service box) the service box itself, conditioned on the presence of a tagged customer, then becomes a M/G/1 system with average arrival rate λ' and service distribution $B(x)$. Therefore, $V^*(t,s)$ can be obtained readily from previous results for M/G/1 system with λ' replacing λ. In the case of FCFS, from Eq. (3.18), we can write down

$$V^*(t,s) = \frac{s(1 - \lambda'/\mu)}{\lambda' B^*(s) - \lambda' + s} \tag{3.19}$$

In order to calculate λ', we refer to Figure 3-2 following Kleinrock [26].



CALCULATION OF THE CONDITIONAL ARRIVAL RATE TO THE SERVICE BOX

FIGURE 3-2

In this figure, assume that two successive customers arrive at time t_1 and t_2; the average time between t_1 and t_2 is clearly $1/\lambda$. Let us also assume that these two customers enter the service box at time t_3 and time t_4, respectively, it is obvious that the average distance between t_3 and t_4 (which is equal to $1/\lambda'$) is larger than that between t_1 and t_2 because the customers in the service box increased their priority at a rate β and the newcomers catch up with them at a rate α as shown by Figure 3-2. In order to calculate λ', we express the vertical offset y in two different ways:

$$y = (\frac{1}{\lambda'})\beta \tag{3.20}$$

$$y = (\frac{1}{\lambda'} - \frac{1}{\lambda})\alpha \tag{3.21}$$

and so λ' is solved as

$$\lambda' = \lambda(1 - \frac{\beta}{\alpha}) \tag{3.22}$$

for convenience, we now define

$$\rho' = \frac{\lambda'}{\mu} = \rho(1 - \frac{\beta}{\alpha}) \tag{3.23}$$

Before we proceed to find $W^*(t,s)$ and $Q^*(s)$, we wish to establish the independent relation between $q(t)$ and $v(t)$ for the SSA system. That is, we wish to prove that the time a customer spends in the queue box is independent of the time he spends (or the time he wastes) in the service box. If this is true, we can find $V^*(t,s)$ and $Q^*(s)$ independently and then multiply them together to get $W^*(t,s)$ (i.e., for two independent random variables, the Laplace transform of the density function of their sum is the product of the individual Laplace transforms). The task will be greatly simplified since $V^*(t,s)$ is available to us already; all we have to do then is to find $Q^*(s)$ from

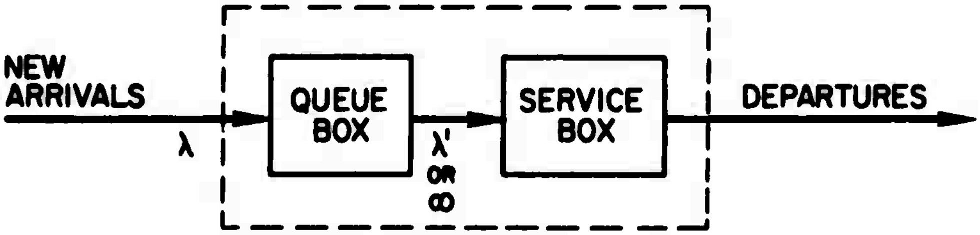
$$Q^*(s) = \frac{W^*(t,s)}{V^*(t,s)} \tag{3.24}$$

Theorem 3.2

For any customer requiring t seconds of service, the time he spends in the queue box is independent of the time he spends in the service box (or independent of the time he wasted in the service box because t is not a random variable).

Proof: See Appendix A

By virtue of Theorem 3.2, all we have to do now is to find $Q^*(s)$, the Laplace transform of the probability density function of the waiting time spent in the queue box. Before we proceed, let us make the following observation:



DECOMPOSITION OF THE SSA SYSTEM

FIGURE 3-3

Figure 3-3 is a modification of Figure 3-1. Arrivals come into the system as a Poisson stream with mean arrival rate λ. If the service box is not idle, customers leave the queue box for the service box at a rate λ'. If the service box becomes idle, then the customer with the highest priority in the queue box (if it is not empty) enters immediately into the service box. This flow of customers keeps on going independent of the scheduling algorithm being used in the service box. From the viewpoint of the queue box, it does not make any difference whether an FCFS, an RR, or any other scheduling algorithm for this matter is being used in the service box. As far as the flow of customers from the queue box to the service box is concerned, the rate is always λ' if the service box is not idle and infinite if it is and if the queue box is not empty. For different scheduling algorithms, there will be different $W^*(t,s)$'s and $V^*(t,s)$'s, but their ratio $Q^*(s)$ remains the same, and it is this $Q^*(s)$ that we are trying to solve (see Eq. 3.24).

Theorem 3.3

The Laplace transform $Q^*(s)$ of the density function of the waiting time spent in the queue box by a customer requiring t seconds of service time (actually, it is independent of t as we explained earlier) is

$$Q^*(s) = \frac{(1-\rho)}{(1-\rho')} \; \frac{\lambda' B^*(s) - \lambda' + s}{\lambda B^*(s) - \lambda + s} \qquad (3.25)$$

with first moment equal to

$$W_q = \lim_{s \to 0} - \frac{\partial Q^*(s)}{\partial s} = \frac{\lambda \overline{t^2}}{2(1-\rho)} - \frac{\lambda' \overline{t^2}}{2(1-\rho')} \qquad (3.26)$$

Proof: See Appendix A

By combining Theorem 3.2 and Theorem 3.3, we get the following important result.

Theorem 3.4

The Laplace transform of the equilibrium waiting time distribution function can be expressed as

$$W^*(t,s) = Q^*(s)\ V^*(t,s)$$

$$= \frac{(1 - \rho)}{(1 - \rho')} \cdot \frac{\lambda' B^*(s) - \lambda' + s}{\lambda B^*(s) - \lambda + s} \cdot V^*(t,s) \tag{3.27}$$

and the mean waiting time as

$$W(t) = W_q + V(t)$$

$$= \frac{\lambda \overline{t^2}}{2(1 - \rho)} - \frac{\lambda' \overline{t^2}}{2(1 - \rho')} + V(t) \tag{3.28}$$

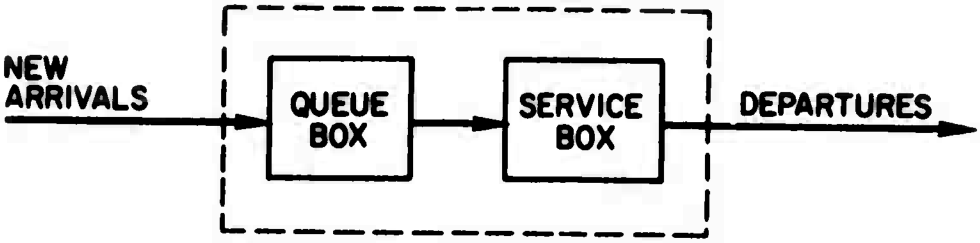
where

$$\overline{t^2} = \int_0^\infty t^2 dB(t) \tag{3.29}$$

Proof: The proof is obvious from the fact that the Laplace transform of the density function of the sum of two independent random variables is just the product of the individual Laplace transforms.

3.3 The Selfish Round Robin (SRR) System

We choose the SRR system as our first example. Figure 3-1 is replotted below to demonstrate the system behavior. Customers arrive as a Poisson stream with average arrival rate λ. Upon their arrival, they enter the queue box where their priority will increase from zero at a positive rate α. After a customer's priority catches up to that of those in the service box, he will join them there and start to share the service facility. All customers in the service box share the service facility equally among themselves as in an RR system; at the same time, their priority increases at a positive rate β. The range of α and β of interest to us is when $\alpha \geq \beta \geq 0$.



DECOMPOSITION OF THE SAA SYSTEM

FIGURE 3-1

Unfortunately for the RR system, the Laplace transform of the waiting time distribution has been solved only for the case of exponential service distribution. For general service distributions, only the mean waiting times are available [23,24] as given by:

$$W_{RR}(t) = \frac{\rho t}{1 - \rho} \tag{3.30}$$

Since the service box looks like an RR system with an average arrival rate λ', the average waiting time in the service box for a customer requiring t seconds of service is the same as in Eq. (4.30) with λ' replacing λ, thus

$$V(t) = \frac{\rho' t}{1 - \rho'} \tag{3.31}$$

From Eq. (3.26), we know that the waiting time in the queue box is

$$W_q = \frac{\lambda \overline{t^2}}{2(1 - \rho)} - \frac{\lambda' \overline{t^2}}{2(1 - \rho')} \tag{3.32}$$

thus, for M/G/1, that we can write down [from Eq. (3.28)]

$$W(t) = \frac{\lambda \overline{t^2}}{2(1 - \rho)} - \frac{\lambda' \overline{t^2}}{2(1 - \rho')} + \frac{\rho' t}{1 - \rho'}$$

$$= \frac{\lambda\overline{t^2}}{2(1 - \rho)} - \frac{\rho'\left(t - \frac{\overline{t^2}}{2\bar{t}}\right)}{(1 - \rho')} \tag{3.33}$$

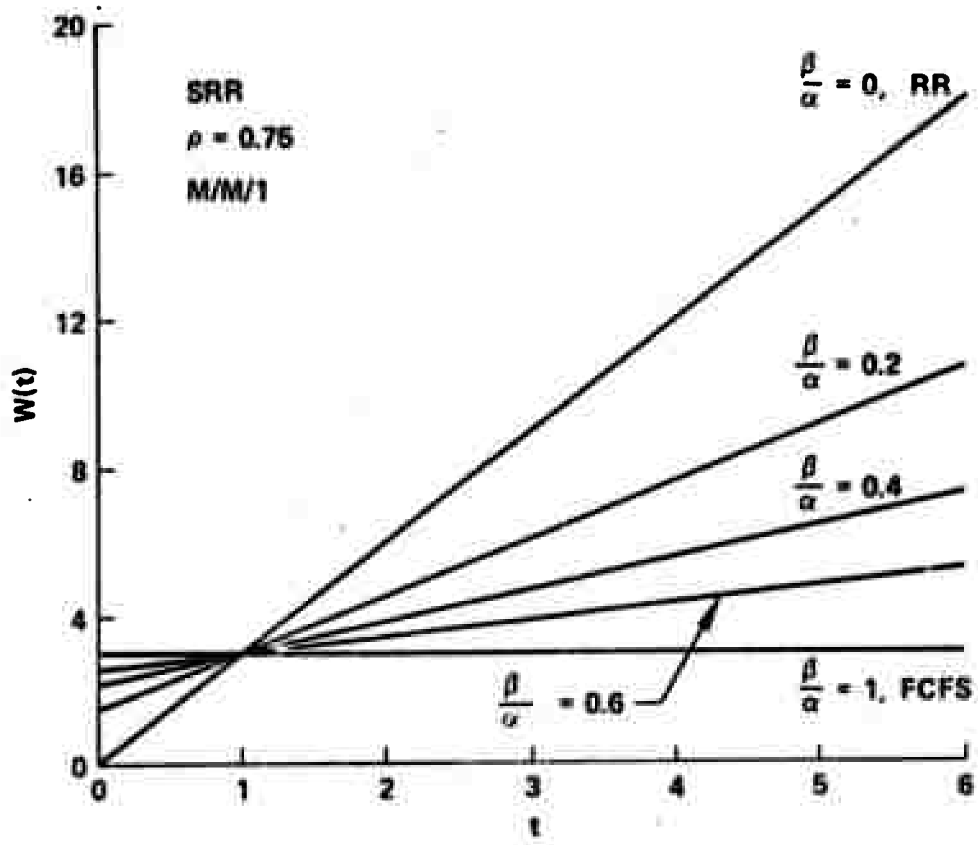
Three examples are given to demonstrate the nature of the mean waiting times for the SRR systems. As our first example, we choose the M/M/1 system (i.e., the service times are exponentially distributed). For this case, the mean waiting time $W(t)$ becomes

$$W(t) = \frac{\rho/\mu}{1 - \rho} - \frac{\rho'/\mu}{1 - \rho'} + \frac{\rho' t}{1 - \rho'}$$

$$= \frac{\rho/\mu}{1 - \rho} + \frac{(t - 1/\mu)\rho'}{1 - \rho'} \tag{3.34}$$

this result was first obtained in [26].

In Figure 3-4, we plot the average waiting time function $W(t)$ against the requested service time t for different ratios of β and α with $\lambda = 0.75$ and $\mu = 1.0$. From Figure 3-4, as well as from Eq. (3.34), we observe that the dependence of $W(t)$ upon t is linear for the entire family of SRR systems; and all the waiting time functions intersect at the same point $(t = \frac{1}{\mu})$. Thus, the performance of a customer who needs exactly $\frac{1}{\mu}$ seconds of service time is the same that he would encounter for any SRR system. In Section 2.2.4, we had observed that correspondance between the RR system and the FCFS system; now we show it holds for the entire class of SRR systems. We also observe that for a customer requesting more than $\frac{1}{\mu}$ seconds of service, his waiting time in the SRR system is longer than that he would experience in the FCFS system; conversely, a customer who requests less than $\frac{1}{\mu}$ seconds of service gets better treatment in the SRR system than in the FCFS system.



AVERAGE WAITING TIME FUNCTIONS FOR THE SRR SYSTEMS WITH EXPONENTIAL SERVICE DISTRIBUTION. $\lambda = 0.75$, $\mu = 1.0$

FIGURE 3-4

We shall show, through the next two examples, that this property holds true for general service distributions as well, although the point of intersections t_i varies for different distributions. It can be easily proved by referring to Eq. (3.33) that $t_i = \frac{\mu \overline{t^2}}{2}$ in general.

As our second example, we choose the system $M/E_2/1$. In this system, we have

$$\frac{dB(x)}{dx} = (2\mu)^2 x\, e^{-2\mu x} \qquad x \geq 0 \tag{3.35}$$

with mean service time equal to $1/\mu$; the second moment of this distribution is $3/2\mu^2$. Figure 3-5 shows the behavior of this system with $\mu = 1$ and $\lambda = 0.75$. Again, the response time curves cross each other at the same point. It can be easily shown that the point of intersection t_i is at

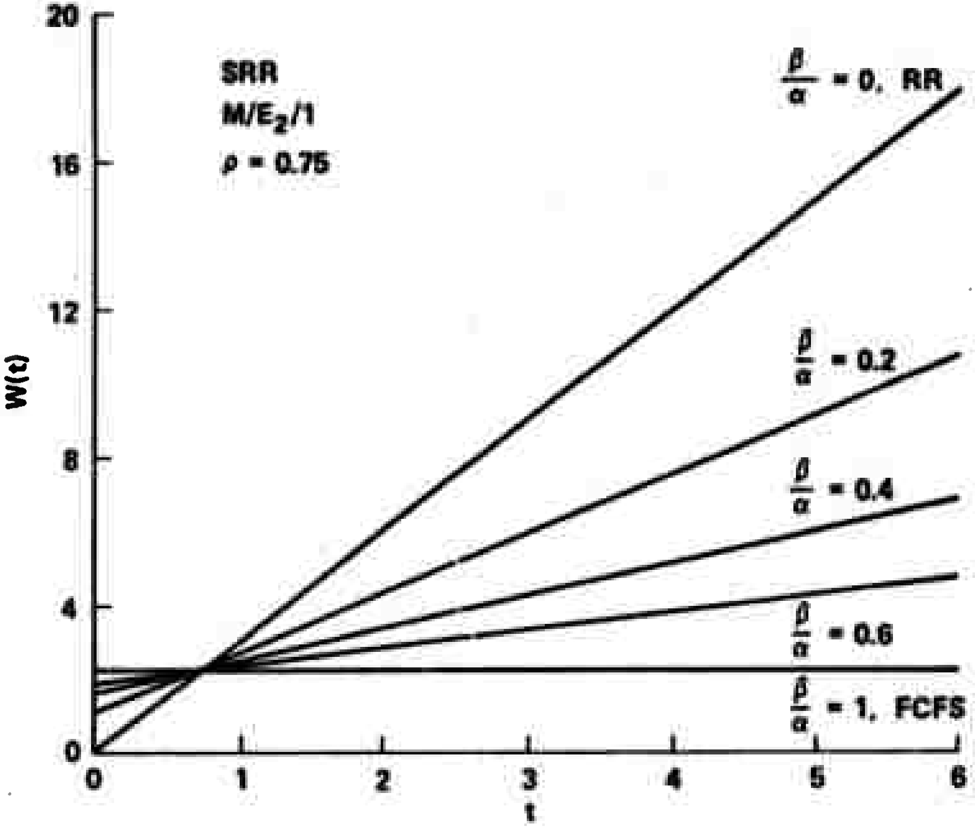
$$t_i = \frac{\mu \overline{t^2}}{2} = 0.75 \tag{3.36}$$

Because the second moment $\overline{t^2}$ is smaller for this distribution than the exponential distribution, t_i is to the left of $1/\mu$.

In the third example, we show the waiting times for the $M/H_2/1$ system, where H_2 stands for hyperexponential service distribution with

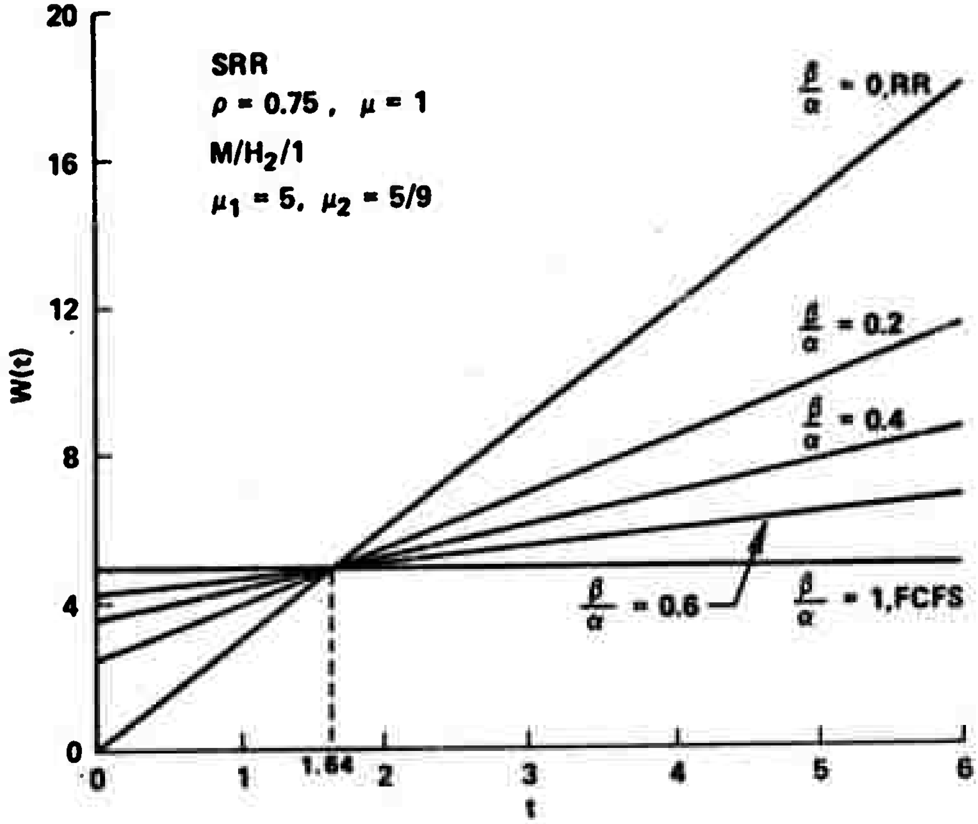
$$\frac{dB(x)}{dx} = 0.5\, \mu_1 e^{-\mu_1 x} + 0.5\, \mu_2 e^{-\mu_2 x} \qquad x \geq 0 \tag{3.37}$$

We choose $\mu_1 = 5\mu$, $\mu_2 = (5/9)\mu$, resulting in a mean service time of $1/\mu$. The second moment of this distribution is $82/25\mu^2$. Figure 3-6 shows the waiting times of the $M/H_2/1$ system with $\mu = 1$ and $\lambda = 0.75$. Again, the waiting times are linearly proportional to t and crossing each other at the same point, only this time, the point of



AVERAGE WAITING TIME FUNCTIONS FOR THE SRR SYSTEMS WITH 2-STAGE ERLANGIAN SERVICE DISTRIBUTION. $\lambda = 0.75$, $\mu = 1.0$

FIGURE 3-5



AVERAGE WAITING TIME FUNCTIONS FOR THE SRR SYSTEMS WITH HYPEREXPONENTIAL SERVICE DISTRIBUTION. $\lambda = 0.75$, $\mu = 1.0$

FIGURE 3-6

intersection t_i is to the right of $1/\mu$ because of the larger second moment of service time distribution.

As we mentioned earlier, the Laplace transform of the waiting time dis ribution $W^*(t,s)$ for the RR systems has been solved only for the M/M/1 case; it is given as [25]

$$W^*_{RR}(t,s) = \frac{(1-\rho)(1-\rho r^2)\, e^{-\lambda(1-r)t}}{(1-\rho r)^2 - \rho(1-r^2)\, e^{-\mu t(1-\rho r^2)/r}} \tag{3.38}$$

where r is taken as the smaller of the two following expressions:

$$r = (\lambda + \mu + s - [(\lambda+\mu+s)^2 - 4\mu\lambda]^{1/2})/2\lambda \tag{3.39}$$

or

$$r = 2\mu\ (\lambda + \mu + s + [(\lambda+\mu+s)^2 - 4\mu\lambda]^{1/2})^{-1} \qquad (4.40$$

From Eq. (3.25), for the M/M/1 system, $Q^*(s)$ is given as

$$Q^*(s) = \frac{(1-\rho)}{(1-\rho')}\ \frac{s+\mu-\lambda'}{s+\mu-\lambda} \tag{3.41}$$

Since $V^*(t,s)$ is the same as $W^*_{RR}(t,s)$ with λ' replacing λ, we can readily write down $W^*(t,s)$ for the M/M/1 SRR system as

$$W^*(t,s) = Q^*(s) \cdot V^*(t,s)$$

$$= \frac{(1-\rho)}{(1-\rho')}\ \frac{s+\mu-\lambda'}{s+\mu-\lambda}\ \frac{(1-\rho')(1-\rho' r'^2)\, e^{-\lambda'(1-r')t}}{(1-\rho' r')^2 - \rho'(1-r'^2)e^{-\mu t(1-\rho' r'^2)/r'}} \tag{3.42}$$

where r' is defined by Eqs. (3.39) and (3.40) with λ replaced by λ'.

By differentiating Eq. (3.42), the first two moments of the waiting time distribution are obtained as, respectively,

$$W(t) = - \frac{\partial W^*(t,s)}{\partial s}|s = 0$$

$$= \frac{\rho/\mu}{1 - \rho} - \frac{\rho'/\mu}{1 - \rho'} + \frac{\rho' \cdot t}{1 - \rho'} \qquad (3.43)$$

$$W_2(t) = \frac{\partial^2 W^*(t,s)}{\partial s^2}|s = 0$$

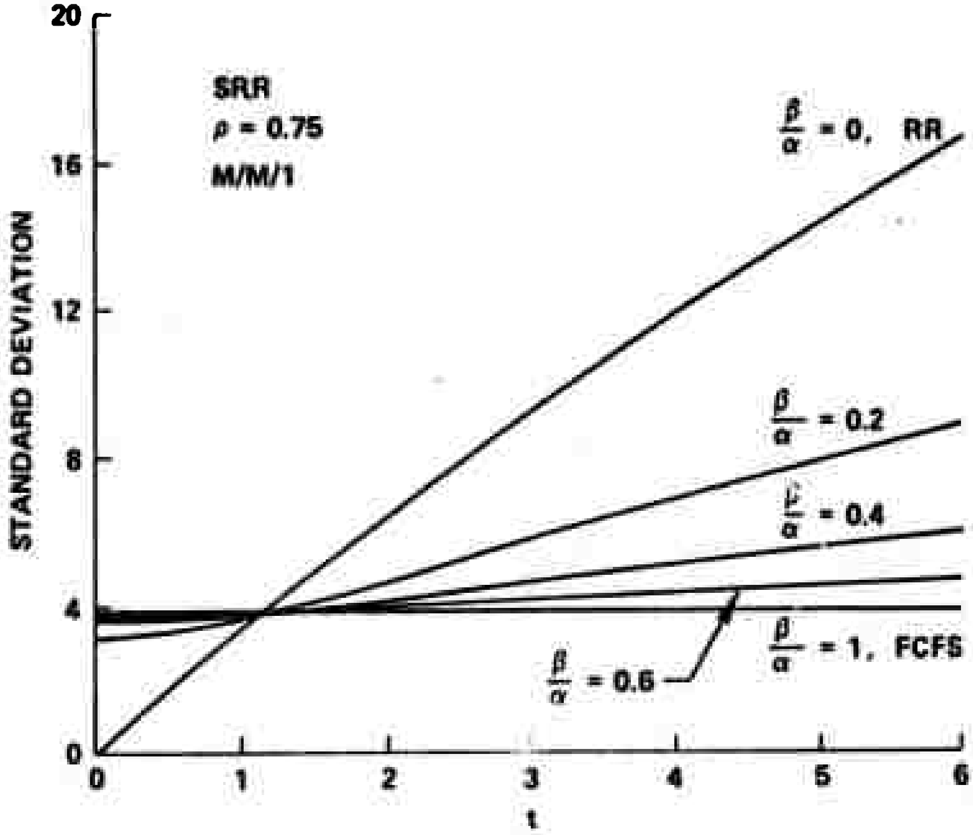
$$= \frac{2\rho' t(\rho - \rho')}{\mu(1 - \rho')^2(1 - \rho)} + \frac{(\rho' t)^2}{(1 - \rho')^2} + \frac{2\rho' t}{\mu(1 - \rho')^3}$$

$$+ \frac{2(\rho - \rho')}{\mu^2(1 - \rho')(1 - \rho)^2} - \frac{2\rho'}{\mu^2(1 - \rho')^4}[1-e^{-(1 - \rho')\mu t}] \qquad (3.44)$$

In Figure 3-7, we plot the standard deviations of the waiting time versus t for different values of β/α. By comparing Figures 3-4 and 3-7, we see that when t is small, the standard deviation tends to be somewhat larger than the mean value; and when t gets large, the mean value tends to be higher than the corresponding standard deviation. In Figure 3-8, we plot the ratio of the standard deviation $\sigma(t)$ to the average of the waiting time against μt with $\lambda = 0.75$ and $\mu = 1.0$. As shown by the figure, $\sigma(t)$ is monotonically nonincreasing with t for the entire family of the SRR systems. It can easily be shown that, when t is large, $\sigma(t)/W(t)$ is proportional to $1/\sqrt{t}$. Thus, the mean waiting times give a better indication of the system behavior when the requested service time is large compared to $1/\mu$.

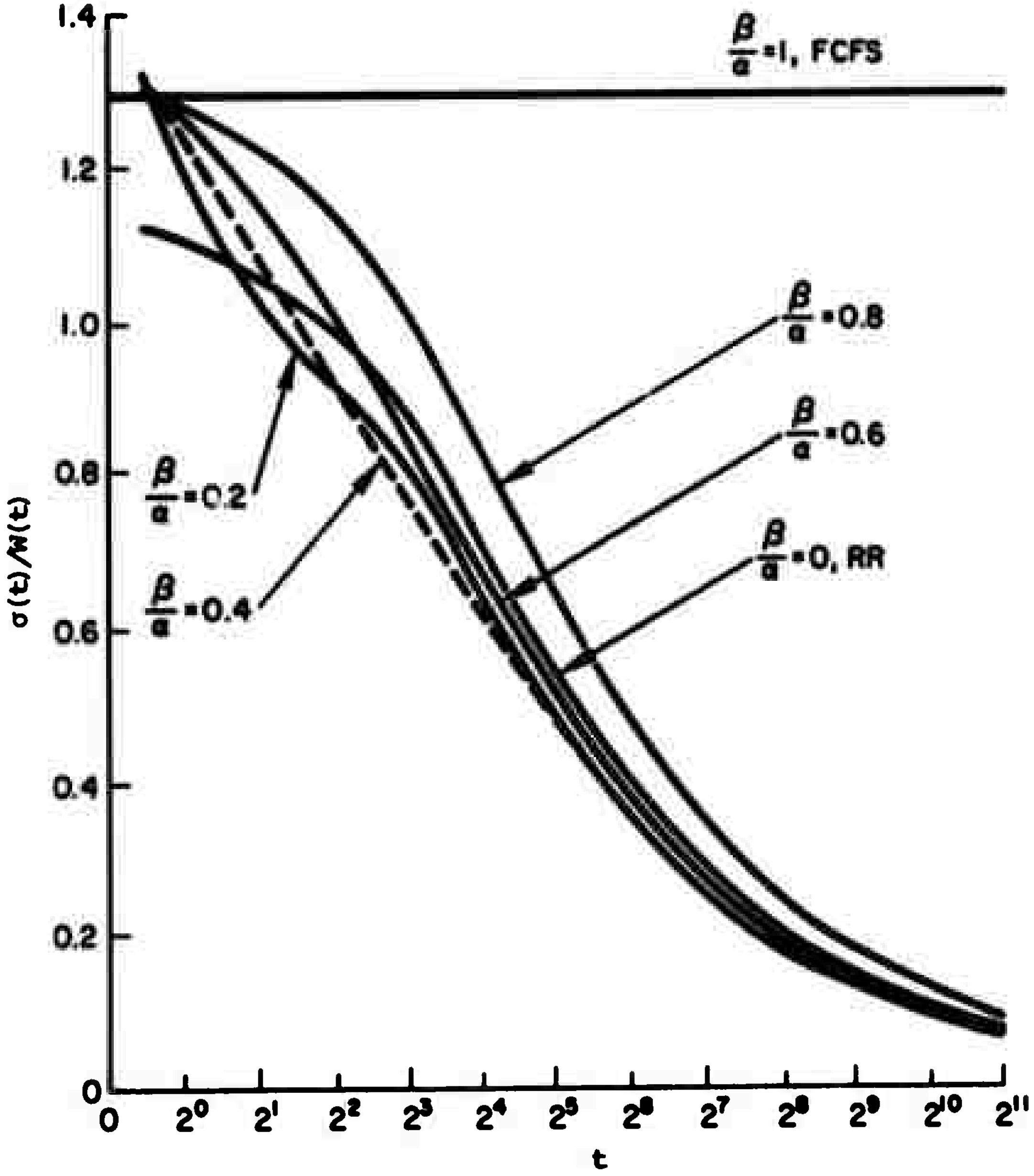
3.4 The Selfish Foreground-Background (SFB) System

The SFB system is very similar to the SRR system we just discussed, the only difference being that the scheduling algorithm being used in the service box is FB instead of RR. Customers enter the service box



STANDARD DEVIATIONS OF THE WAITING TIME FOR THE SRR SYSTEMS WITH EXPONENTIAL SERVICE DISTRIBUTION. $\lambda = 0.75$, $\mu = 1.0$

FIGURE 3-7



σ(t)/W(t) PLOTTED AGAINST μt FOR THE SRR
M/M/1 SYSTEM WITH λ = 0.75, μ = 1.0

FIGURE 3-8

after they have experienced some waiting in the queue box. Once a customer enters the service box, he will occupy the service facility immediately all by himself because he is the one with the least amount of attained service in the service box. From there on, this "tagged" unit sees a pure FB system with a Poisson arrival process at λ' customers per second until he leaves the system completely served.

The Laplace transform of the response time distribution for FB systems with general service time distributions has been solved by Schrage [27]. Thus, by substituting λ with λ' in his results, we get

$$Y^*(t,s) = H^*(t,s) \tag{3.45}$$

where

$$D^*(t,s) = B^*[s + \lambda' \{1 - A^*(t,s)\}] \tag{3.46}$$

$$A^*(t,s) = G^*[t,s + \lambda' \{1 - A^*(t,s)\}] \tag{3.47}$$

$$H^*(t,s) = (1 - \rho') \; [s + \lambda' \{1 - A^*(t,s)\}] \tag{3.48}$$

and $G^*(t,s)$ is defined as

$$\lim_{s\to 0} (-1)^m \frac{\partial^m G^*(t,s)}{\partial s^m} = \int_0^t t^m dB(t) + t^m[1 - B(t)] \tag{3.49}$$

Substitute Eq. (3.45) into Eq. (3.27) and we get

$$S^*(t,s) = \frac{1 - \rho}{1 - \rho'} \; \frac{\lambda' B^*(s) - \lambda' + s}{\lambda B^*(s) - \lambda + s} \; H^*(t,s) \; D^*(t,s) \tag{3.50}$$

From Eq. (3.50), we can derive the first two moments of the waiting time distribution of the SFB system as

$$W(t) = \frac{\lambda \overline{t^2}}{2(1 - \rho)} - \frac{\lambda' \overline{t^2}}{2(1 - \rho')} + \frac{\lambda' \overline{t^2}_{<t}}{2(1 - \rho'_{<t})^2} + \frac{t(\rho'_{<t})}{1 - \rho'_{<t}} \tag{3.51}$$

$$W(t^2) = \frac{\lambda' \overline{t^3}_{<t}}{3(1 - \rho'_{<t})^3} + \frac{(\lambda' \overline{t^2}_{<t})^2}{(1 - \rho'_{<t})^4} + \frac{t^2(\rho'_{<t})^2}{(1 - \rho'_{<t})^2} + \frac{\lambda' t \cdot \overline{t^2}_{<t}}{(1 - \rho'_{<t})^3}$$

$$+ \frac{2(\rho - \rho')}{\mu^2(1 - \rho')(1 - \rho)^2} + 2\frac{(\rho - \rho')}{\mu(1 - \rho)(1 - \rho')} \left[\frac{\lambda' \overline{t^2}_{<t}}{2(1 - \rho'_{<t})^2} + \frac{t \cdot \rho'_{<t}}{1 - \rho'_{<t}}\right] \tag{3.52}$$

where

$$\overline{t^m}_{<t} = \int_0^t t^m dB(t) + t^m[1 - B(t)] \tag{3.53}$$

and

$$\rho'_{<t} = \lambda' \cdot \bar{t}_{<t} \tag{3.54}$$

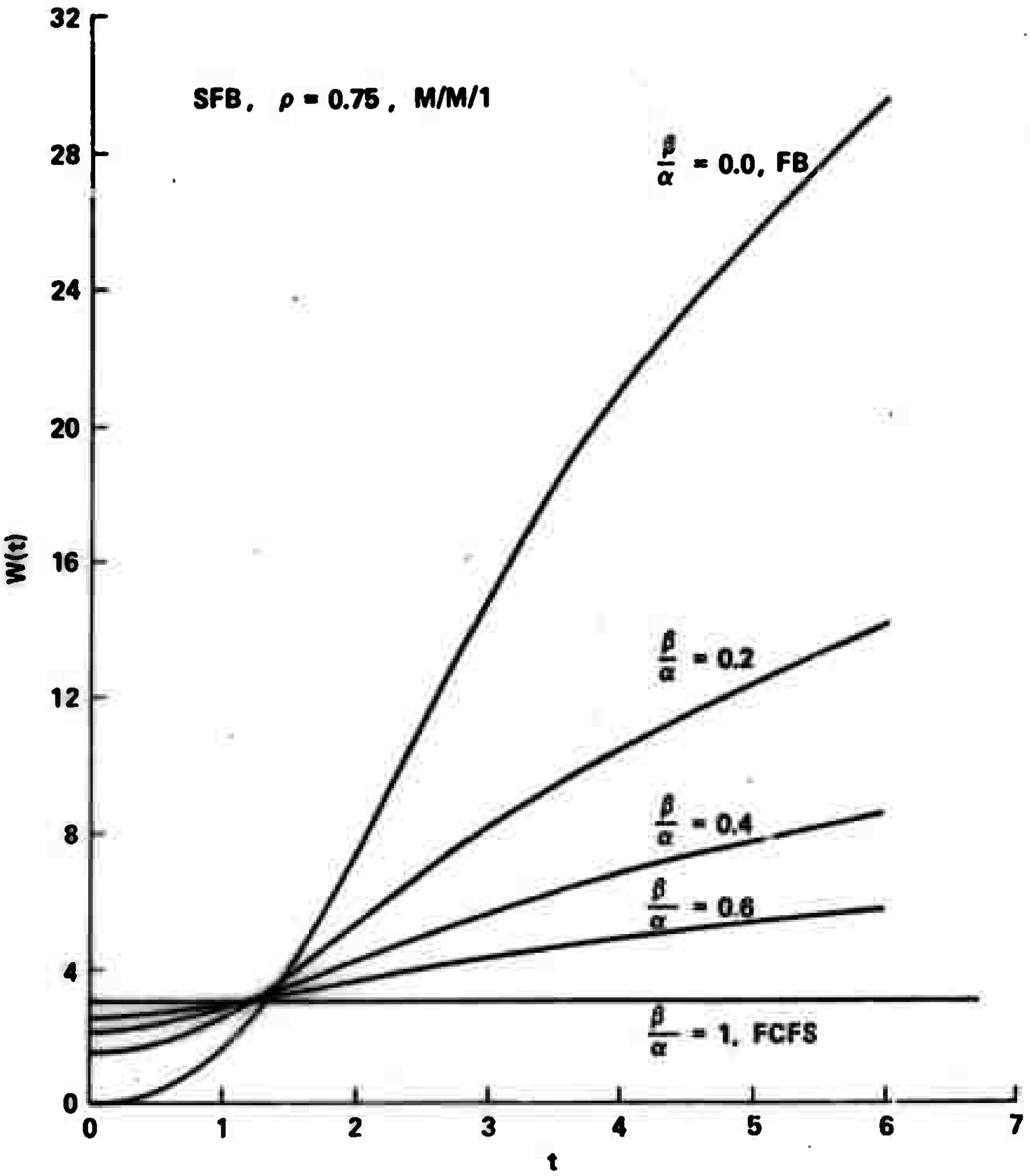
Three examples are given to show the nature of the waiting time we just derived. As the first example, we again choose the M/M/1 system. In Figure 3-9, waiting times for different β/α ratios are plotted against t with the assumption that $\lambda = 0.75$, $\mu = 1$. When $\beta/\alpha = 0$, the system becomes a pure FB system with no waiting in the queue box. As the ratio β/α increases, the average wait in the queue box also increases until it hits the maximum point when $\beta/\alpha = 1$, which happens to be a pure FCFS system. The curves representing different β/α ratios do not intersect the same point as in the SRR system, but the points of intersections are relatively close to each other. Figure 3-10 shows the standard deviations as plotted against t. In Figure 3-11, $\sigma(t)/W(t)$ is plotted against μt with $\lambda = 0.75$ and $\mu = 1$. When t is small, the standard deviation tends to be somewhat larger than the mean waiting time, indicating a rather large zone for the mean value to vary. When t gets large, the ratio $\sigma(t)/W(t)$ varies with $1/\sqrt{t}$, the standard deviation levels off faster than does the mean value, thus making the mean waiting time a "better" result.

As our second example, we choose the system $M/E_2/1$. The service time distribution is defined as Eq. (3.35). Figure 3-12 shows the behavior of this system with $\mu = 1$ and $\lambda = 0.75$. These figures are somewhat similar to those for the M/M/1 system, with smaller average waiting time in the queue box. Figure 3-13 shows the standard deviations as plotted against t. For β/α not equal to zero, the standard deviation assumes a rather large value at $t = 0$, then tapers off a bit before going up again. When t is a few times larger than the average service time $1/\mu$, the mean waiting times are much higher than their corresponding standard deviations, indicating that relatively small range where the waiting times can fall.

Once more, we choose the $M/H_2/1$ system defined by Eq. (3.37) as our third example. With $\lambda = 0.75$ and $\mu = 1$, mean waiting times and standard deviations are plotted against t as shown in Figures 3-14 and 3-15, respectively. The average waiting times in the queue box (as indicated at $t = 0$ in Figure 3-14) are larger than their corresponding terms in an M/M/1 system because the second moment of the hyperexponential service distribution is larger. Again, when t is large compared to $1/\mu$, the standard deviations tend to be less than their corresponding mean waiting times in a similar manner as were mentioned earlier for the M/M/1 and $M/E_2/1$ systems.

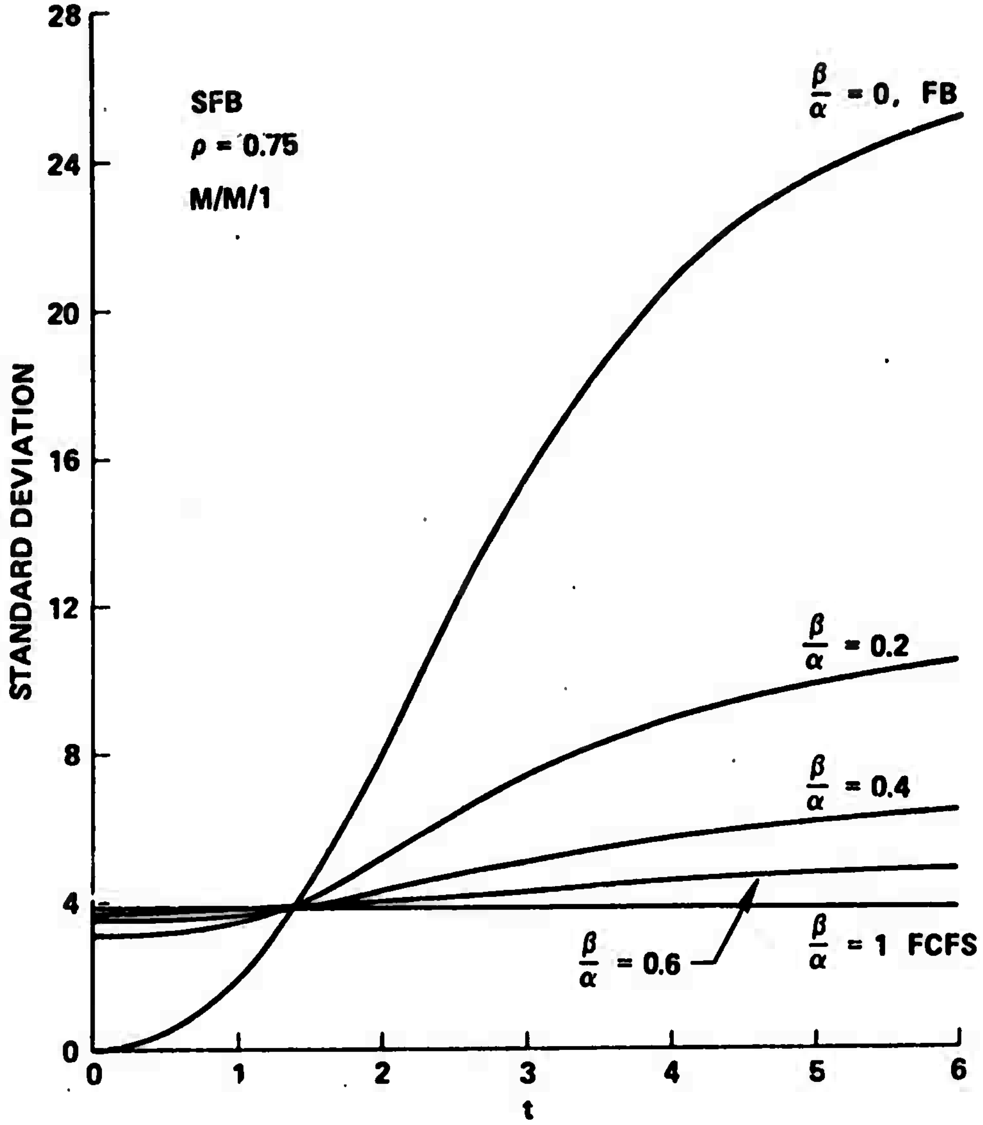
3.5 Summary

In this chapter, we discussed the family of selfish scheduling algorithms. The results obtained in Section 3.2 can be applied to any



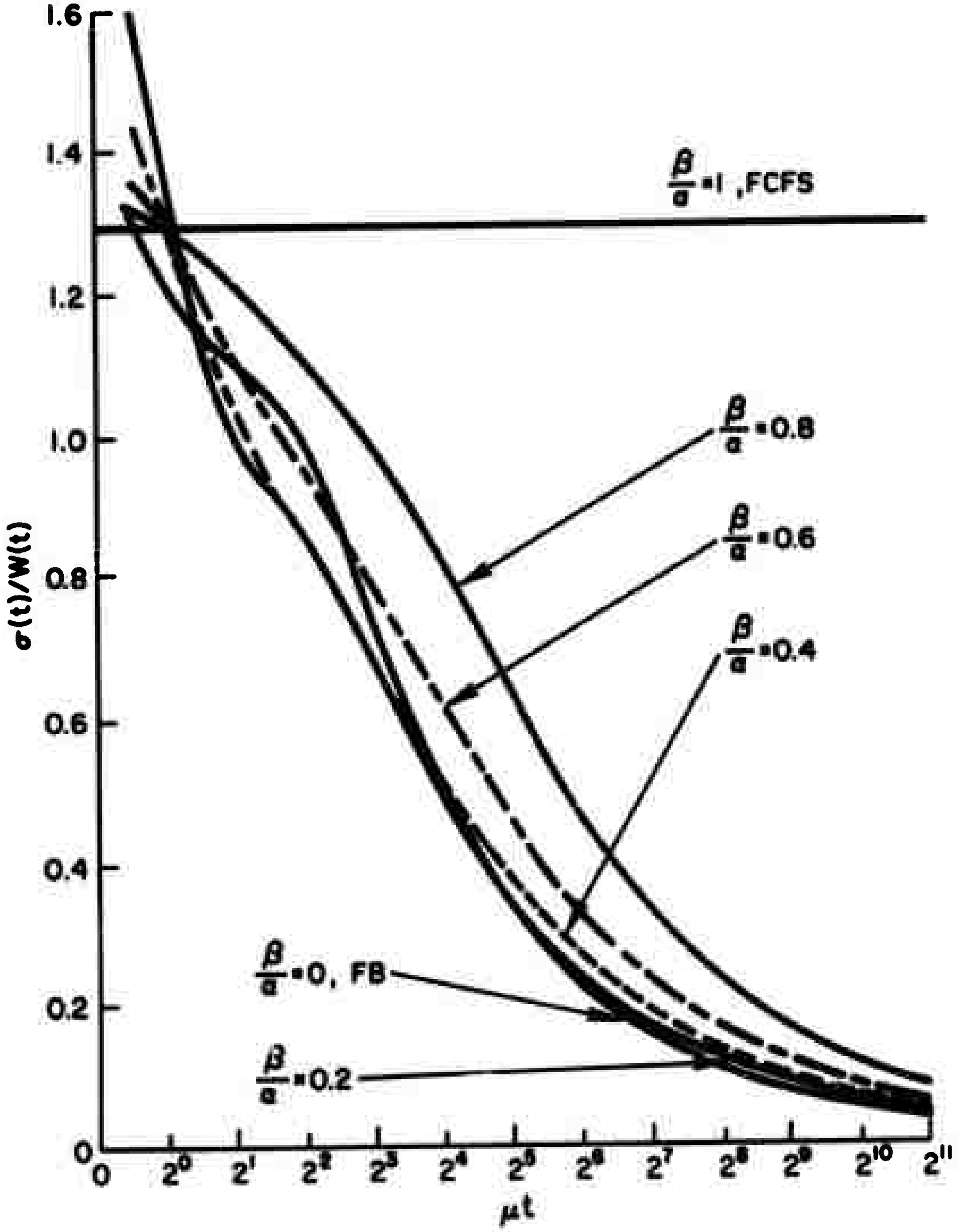
AVERAGE WAITING TIME FUNCTIONS FOR THE SFB SYSTEMS WITH EXPONENTIAL SERVICE DISTRIBUTION. $\lambda = 0.75$, $\mu = 1.0$

FIGURE 3-9



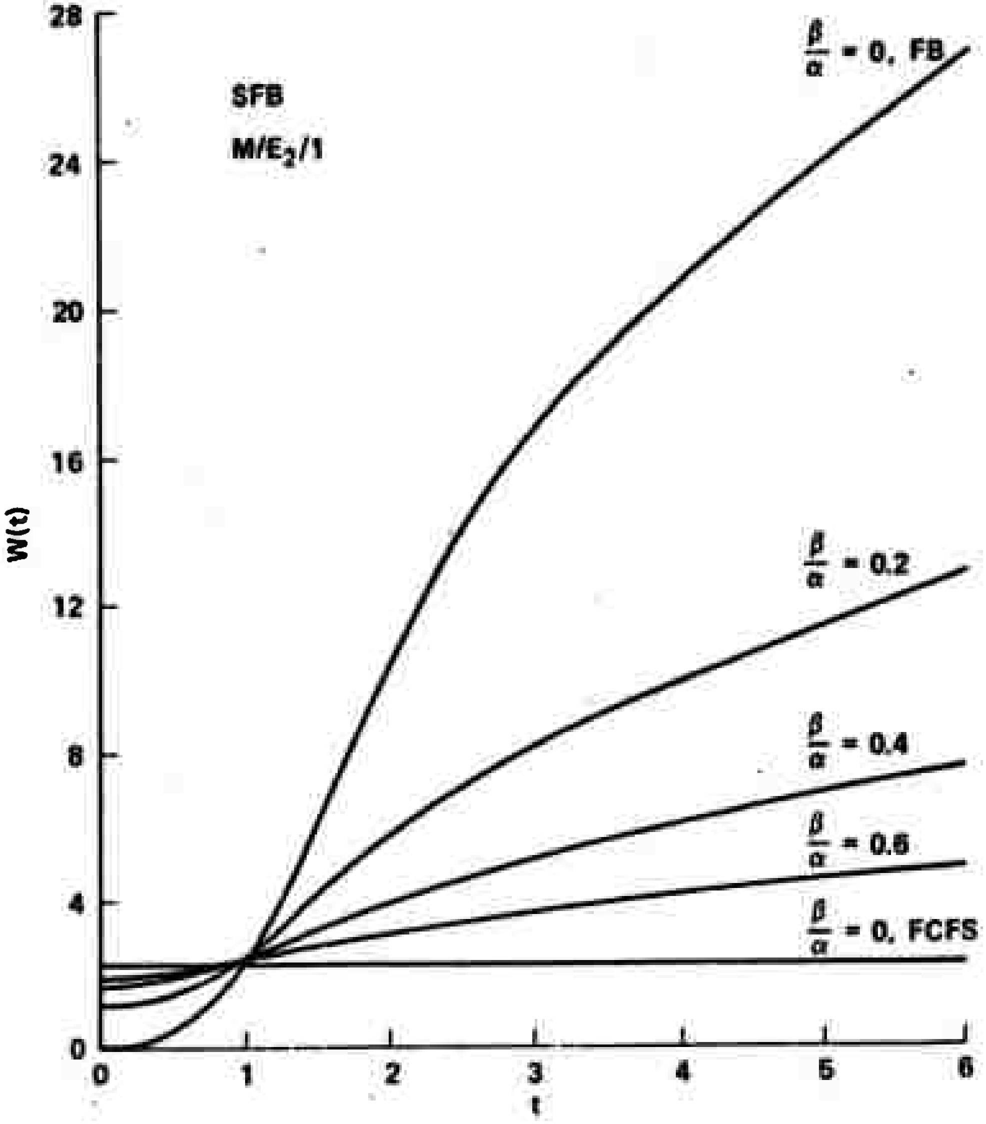
STANDARD DEVIATIONS OF THE WAITING TIME FOR THE SFB SYSTEMS WITH EXPONENTIAL SERVICE DISTRIBUTION. $\lambda = 0.75$, $\mu = 1.0$

FIGURE 3-10



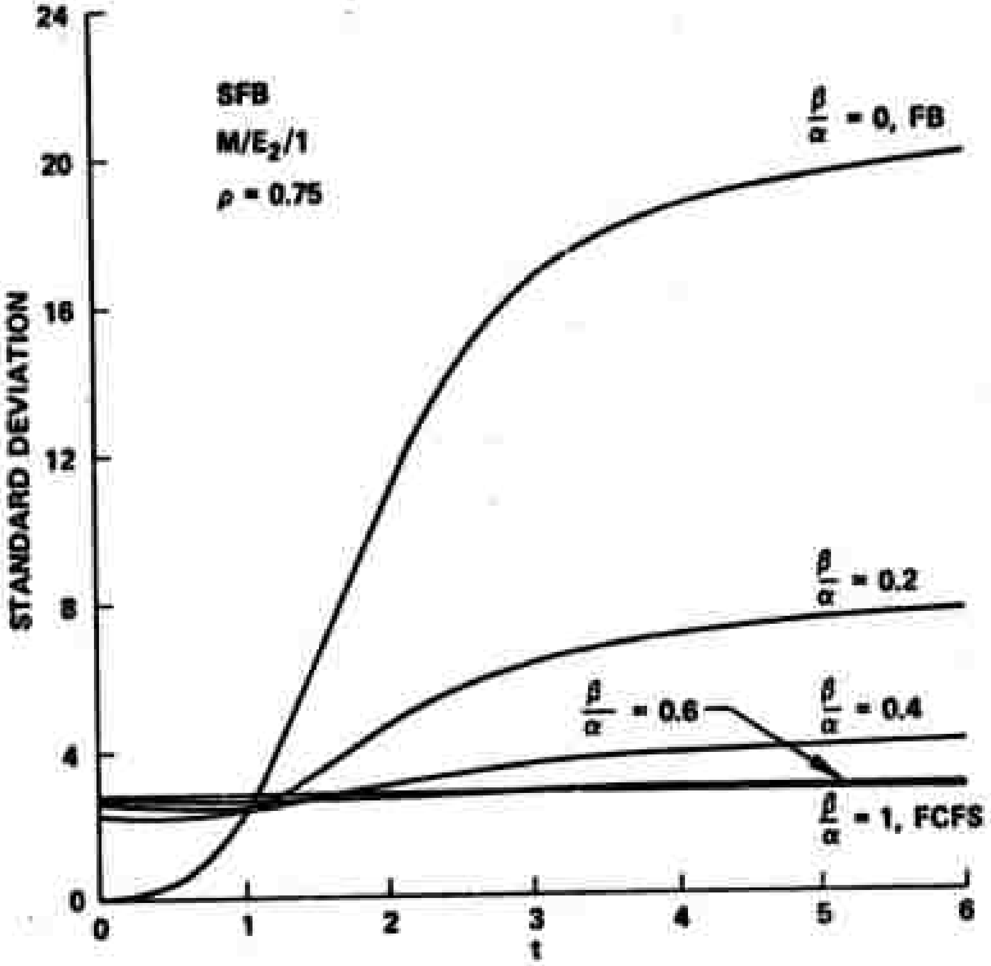
σ(t)/W(t) PLOTTED AGAINST t FOR SFB M/M/1
SYSTEMS WITH λ = 0.75 AND μ = 1.0

FIGURE 3-11



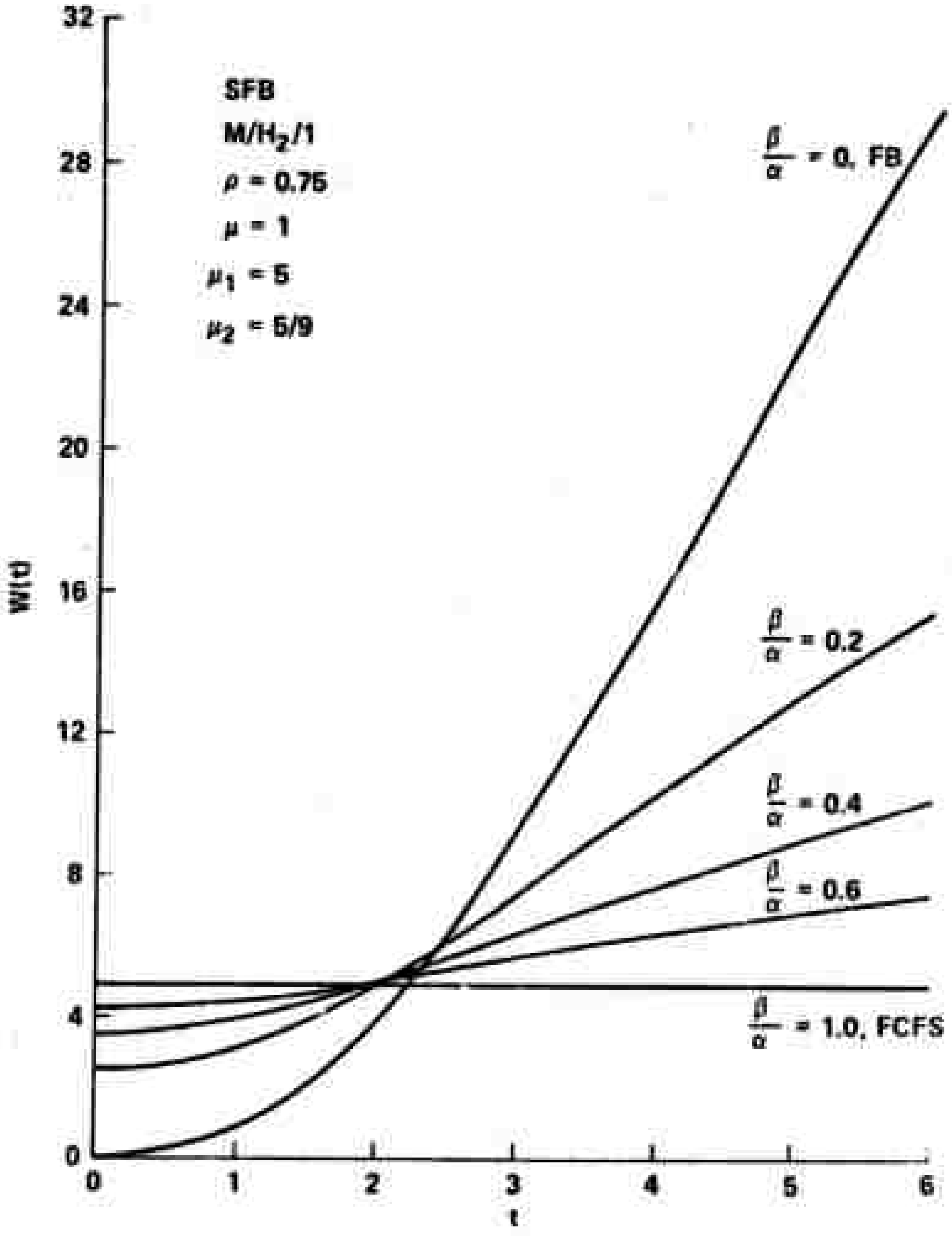
AVERAGE WAITING TIME FUNCTIONS FOR THE SFB SYSTEMS WITH 2-STAGE ERLANGIAN SERVICE DISTRIBUTION. $\lambda = 0.75$, $\mu = 1.0$

FIGURE 3-12



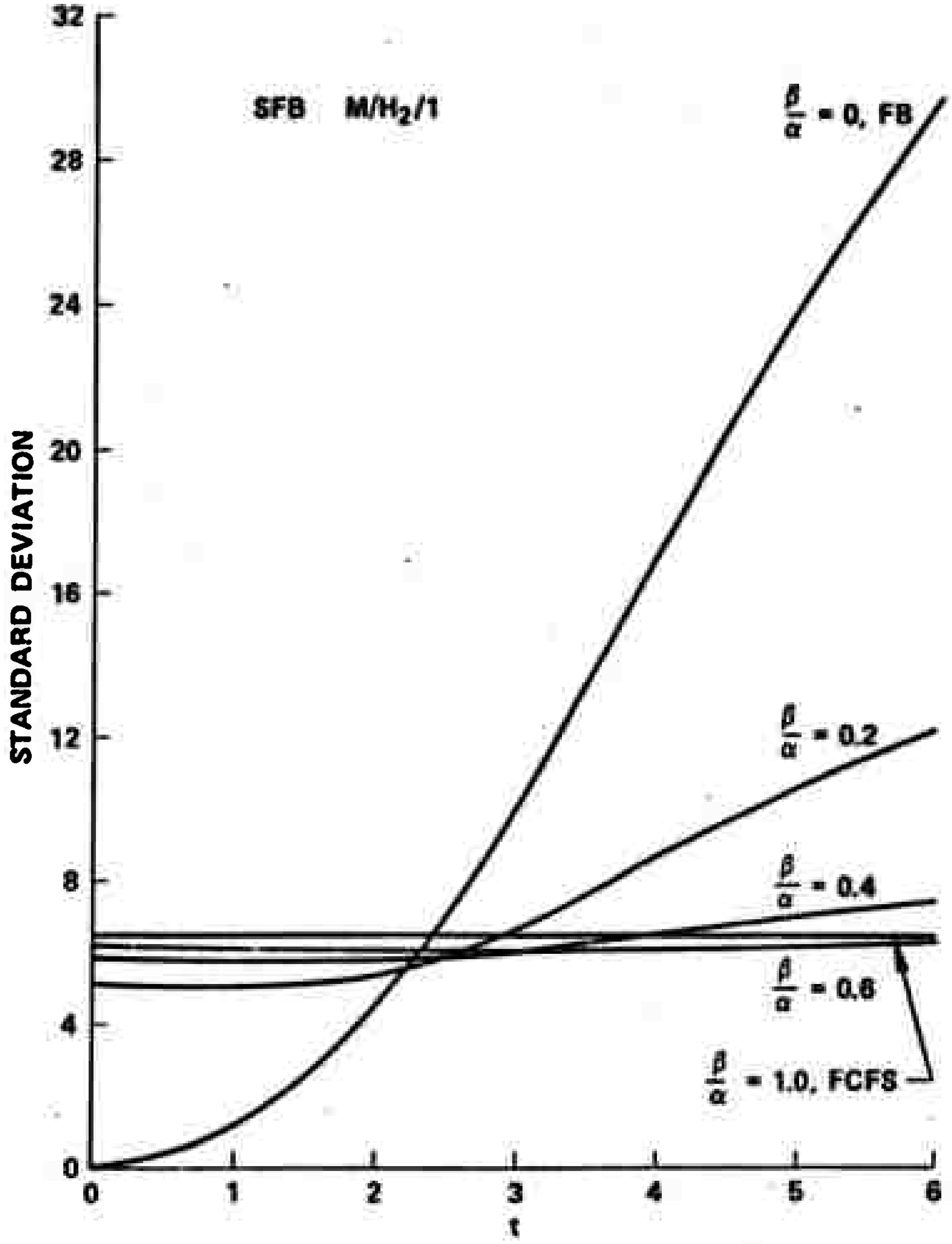
STANDARD DEVIATIONS OF THE WAITING TIME FOR THE SFB SYSTEMS WITH 2-STAGE ERLANGIAN SERVICE DISTRIBUTION. $\lambda = 0.75$, $\mu = 1.0$

FIGURE 3-13



AVERAGE WAITING TIME FUNCTIONS FOR THE SFB SYSTEMS WITH HYPEREXPONENTIAL SERVICE DISTRIBUTION. $\lambda = 0.75$, $\mu = 1.0$

FIGURE 3-14



STANDARD DEVIATIONS OF THE WAITING TIME FOR THE SFB SYSTEMS WITH HYPEREXPONENTIAL SERVICE DISTRIBUTION. $\lambda = 0.75$, $\mu = 1.0$

FIGURE 3-15

SSA system. Two parameters, α and β, are introduced to the system so that one degree of freedom (appearing as the ratio β/α) is provided to the system designer which he can use to adjust the system performance as a function of service time over a continuum of service behaviors.

Two specific systems, the selfish round-robin and the selfish foreground-background, are described in detail to demonstrate the nature of the results. For the SRR system, the average response time always varies linearly with the requested service time; the Laplace transform of the waiting time distribution is available only for the exponential service time distribution. When the SFB algorithm is used by the system scheduler, the Laplace transform of the response time distribution is obtained for M/G/1 system in general. Examples for exponential, Erlangian, and hyperexponential service distribution are given in Sections 3.3 and 3.4.

Once again we wish to emphasize the generality of the results obtained in this chapter. Given the result of any M/G/1 system,time-shared system analysis, the result for the corresponding selfish system can be readily obtained by referring to Eq. (3.27) (Laplace transform of the waiting time) and Eq. (3.28) (the average waiting time).

CHAPTER 4

A CONTINUUM OF FEEDBACK SCHEDULING ALGORITHMS

4.1 The Mathematical Model

In Chapter 3 we discussed the family of selfish scheduling algorithms. In particular, we discussed the SRR and the SFB systems where the performance of those systems varies over a continuum of system behaviors with the FCFS system at one end of the continuum (i.e. $\beta/\alpha = 1$). The RR system lies at the other end of the continuum (i.e. $\beta/\alpha = 0$) for the SRR systems; and the FB system lies at this end ($\beta/\alpha = 0$) for the family of SFB systems. In this chapter we look into another family of scheduling algorithms whose performance also ranges over a continuum of system behaviors.

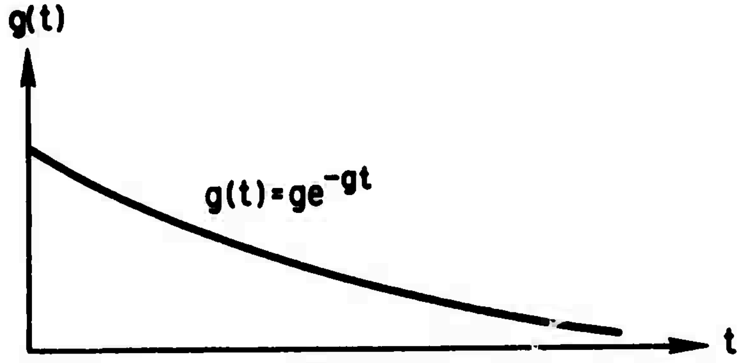
In Section 2.2.5 we discussed the RR system with externally assigned priorities. In such a system, we assume that there are P priority groups each with Poisson arrival process, and an exponentially distributed service request for customers from each group. A positive number g_p is associated with the p^{th} priority group, with larger values of g_p being given to those higher priority groups. All g_p's are assumed to be of constant values. g_p can be interpreted as the relative fraction of the total service facility (the CPU time) that is allocated for the customers from the p^{th} priority group.

The model we introduce in this chapter is an extension of this Round Robin system with priorities. The g_p's are no longer assigned externally, rather they are assigned to the customers according to their

Preceding page blank

attained services t (which also implies that we transform the discrete priority system into a continuous priority system). In particular, we let g(t) vary exponentially with t, namely,

$$g(t) = ge^{-gt} \tag{4.1}$$



g AS FUNCTION OF ATTAINED SERVICE TIME

FIGURE 4-1

As shown by Figure 4-1 and Eq. (4.1), all customers in the system share the service facility simultaneously, and customers with lesser amounts of attained service get served at a higher rate than those with greater attained services. The more service a customer has received, the slower he will get served. Let n(t) denote the density of the number of customers in the system with t seconds of attained service (Section 2.2.9). Since his rate of attaining service is directly proportional to g(t), the fraction $\tilde{f}(t)$ (a random variable) of the total service facility allocated to a customer with t seconds of attained service is calculated as

$$\tilde{f}(t) = \frac{g(t)}{\int_0^\infty g(t)n(t)dt} = \frac{e^{-gt}}{c} \tag{4.2}$$

where c is a constant independent of t.

The RR system is a special case of this family of algorithms. With very small value of g, the slope of the exponential curve drops very slowly with t, and so the discrimination against customers with large attained services (as measured by the relative rates at which they accumulate service time) decreases as g decreases. As g goes to zero, the exponential curve in Figure 4-1 becomes a horizontal (non-discriminatory) line, and every customer get treated equally at all times (independent of his attained service time). This, of course, becomes the Round Robin system.

On the other hand, if we let the value of g be very large, the curvature of g(t) becomes very steep. This means that a customer with a slightly lesser amount of accumulated service time will be served at a much higher rate than another customer in the system with slightly more attained service. The higher the value of g, the more pronounced is this discrimination against long jobs. As g goes to infinity, the only time a customer can get into the service facility is when he is the customer with the least amount of attained service in the system, thus the system becomes FB.

The parameter g can assume any value between 0 and infinity. By varying g, a degree of freedom is provided to the system designer which he may use to adjust the system performance over a continuum of behaviors with the RR and the FB systems serving as the two boundaries.

4.2 The Analysis

We consider the case M/M/1 with scheduling algorithm g(t) as defined above.

Theorem 4.1

The average response time $T(t)$ for the system with the scheduling algorithm defined by $g(t) = ge^{-gt}$, is the solution of the following integro-differential equation

$$(1 - \rho)T(t) = t + \frac{\rho/\mu}{1 - \rho} - \rho\int_0^\infty T'(\tau)\,[e^{g\tau} + e^{gt} - 1]^{-\mu/g} d\tau$$

$$- \rho\int_0^t T'(\tau)\,[e^{gt} - e^{g\tau} + 1]^{-\mu/g} d\tau \qquad (4.3)$$

Proof: See Appendix B

In Eq. (4.3), if $g = 0$, it gives us the result of the RR system. As $g \to 0$, the term $[e^{g\tau} + e^{gt} - 1]^{-\mu/g}$ becomes

$$\lim_{g\to 0} [e^{g\tau} + e^{gt} - 1]^{-\mu/g} = \lim_{g\to 0}\left\{[1 + g\tau + 1 + gt - 1]^{1/g}\right\}^{-\mu}$$

$$= \lim_{g\to 0}\left\{[1 + g(t + \tau)]^{1/g}\right\}^{-\mu}$$

$$= \left[e^{t + \tau}\right]^{-\mu}$$

$$= e^{-\ (t + \tau)} \qquad (4.4)$$

The limit of $[e^{gt} - e^{g\tau} + 1]^{-\mu/g}$ as g goes to zero can similarly be calculated as

$$\lim_{g\to 0} [e^{gt} - e^{g\tau} + 1]^{-\mu/g} = \lim_{g\to 0}\left\{[1 + gt - 1 - g\tau + 1]^{1/g}\right\}^{-\mu}$$

$$= \lim_{g\to 0}\left\{[1 + g(t - \tau)]^{1/g}\right\}^{-\mu}$$

$$= \left[e^{(t - \tau)}\right]^{-\mu}$$

$$= e^{-\mu(t - \tau)} \qquad (4.5)$$

Substituting Eqs. (4.4) and (4.5) into Eq. (4.3), we get

$$(1 - \rho)T(t) = t + \frac{\rho/\mu}{1 - \rho} - \rho\int_0^{\infty} T'(t)e^{-\mu(t + \tau)}d\tau - \rho\int_0^{t} T'(t)e^{-\mu(t - \tau)}d\tau$$

$$= t + \frac{\rho/\mu}{1 - \rho} - \rho e^{-\mu t}\int_0^{\infty} T'(\tau)e^{-\mu\tau}d\tau - \rho e^{-\mu t}\int_0^{t} T'(\tau)e^{\mu\tau}dt$$

$$(4.6)$$

Since from [36], we have

$$\int_0^{\infty} T'(\tau)e^{-\mu\tau}d\tau = \frac{1/\mu}{1 - \rho} \qquad (4.7)$$

Eq. (4.6) becomes

$$(1 - \rho)T(t) = t + \frac{\rho/\mu}{1 - \rho} - \frac{\rho/\mu}{1 - \rho}e^{-\mu t} - \rho e^{-\mu t}\int_0^{t} T'(\tau)e^{\mu\tau}d\tau \qquad (4.8)$$

Multiply $e^{\mu t}$ on both sides of Eq. (4.8) to yield

$$(1 - \rho)e^{\mu t}T(t) = te^{\mu t} + \frac{\rho/\mu}{1 - \rho}(e^{\mu t} - 1) - \rho\int_0^{t} T'(\tau)e^{\mu\tau}d\tau \qquad (4.9)$$

Now differentiate Eq. (4.9) with respect to t, to get

$$(1 - \rho)^{\mu t}T'(t) + \mu(1 - \rho)T(t)e^{\mu t}$$

$$= (1 + \mu t)e^{\mu t} + \frac{\rho}{1 - \rho}e^{\mu t} - \rho e^{\mu t}T'(t) \qquad (4.10)$$

Multiplying $e^{-\mu t}$ on both sides of Eq. (4.10) and simplifying, we get

$$T'(t) + \mu(1 - \rho)T(t) = \frac{1}{1 - \rho} + \mu t \qquad (4.11)$$

With the initial condition $T(0) = 0$, Eq. (4.11) is easily solved as

$$T(t) = \frac{t}{1 - \rho} \tag{4.12}$$

which, of course, is the average response time for the RR system.

On the other hand, when g goes to infinity, for $\left[e^{gt} + e^{g\tau} - 1\right]^{-\mu/g}$, we must consider two cases:

Case 1) when $\tau < t$, then $1 << e^{g\tau} << e^{gt}$

$$\begin{aligned} \lim_{g\to\infty}\left[e^{gt} + e^{g\tau} - 1\right]^{-\mu/g} &= \lim_{g\to\infty}\left[e^{gt}\right]^{-\mu/g} \\ &= e^{-\mu t} \end{aligned} \tag{4.13}$$

Case 2) when $\tau > t$, then $1 << e^{gt} << e^{g}$

$$\begin{aligned} \lim_{g\to\infty}\left[e^{gt} + e^{g\tau} - 1\right]^{-\mu/g} &= \lim_{g\to\infty}\left[e^{g\tau}\right]^{-\mu/g} \\ &= e^{-\mu\tau} \end{aligned} \tag{4.14}$$

Similarly, the limit of $\left[e^{gt} - e^{g\tau} + 1\right]^{-\mu/g}$ as g goes to infinity becomes

$$\begin{aligned} \lim_{\substack{g\to\infty \\ \tau \le t}}\left[e^{gt} - e^{g\tau} + 1\right]^{-\mu/g} &= \lim_{g\to\infty}\left[e^{gt}\right]^{-\mu/g} \\ &= e^{-\mu t} \end{aligned} \tag{4.15}$$

Substituting Eqs. (4.13), (4.14) and (4.15) into Eq. (4.13), we get

$$\begin{aligned} (1 - \rho)T(t) &= t + \frac{\rho/\mu}{1 - \rho} - \int_0^t T'(\tau)e^{-\mu t}d\tau \\ &- \rho\int_t^\infty T'(\tau)e^{-\mu\tau}d\tau - \rho\int_0^t T'(\tau)e^{-\mu t}d\tau \end{aligned} \tag{4.16}$$

or

$$(1 - \rho + 2\rho e^{-\mu t})T(t) = t + \frac{\rho/\mu}{1 - \rho} - \rho\int_t^\infty T'(\tau)e^{-\mu\tau}d\tau \qquad (4.17)$$

Differentiating Eq. (4.17) with respect to t yields

$$T'(t)(1 - \rho + 2\rho e^{-\mu t}) - 2\lambda e^{-\mu t}T(t) = 1 + \rho\, T'(t)e^{-\mu t} \qquad (4.18)$$

Rearranging, we get

$$T'(t)(1 - \rho + \rho e^{-\mu t}) = 2\lambda e^{-\mu t}T(t) + 1 \qquad (4.19)$$

With the initial condition $T(0) = 0$, the solution of Eq. (4.19) is

$$T(t) = \frac{\frac{\rho}{\mu}\left[1 - e^{-\mu t} - \mu t e^{-\mu t}\right]}{(1 - \rho + \rho e^{-\mu t})^2} + \frac{t}{1 - \rho + \rho e^{-\mu t}} \qquad (4.20)$$

which, of ccurse, is the average response time for the FB system (with exponential service time distribution).

Unfortunately, we cannot solve Eq. (4.2) analytically in general except for the special cases with $g = 0$ and $g = \infty$. An approximate solution of Eq. (4.2) with large values of g is presented in the next section.

4.3 An Approximate Solution

When g is large compared to μ, the term $\left|e^{g\tau} + e^{gt} - 1\right|^{-\mu/g}$ can be approximated by using

$$(1 + \varepsilon)^k \cong 1 + \varepsilon k \qquad \text{for } \varepsilon \ll 1 \qquad (4.21)$$

We consider two cases:

Case 1) $\tau < t$

$$\left[e^{gt} + e^{g\tau} - 1\right]^{-\mu/g} = e^{-\mu t}\left[1 + e^{-gt}(e^{g\tau} - 1)\right]^{-\mu/g}$$

$$\simeq e^{-\mu t}[1 - (\mu/g)e^{-gt}(e^{g\tau} - 1)] \qquad (4.22)$$

Case 2) $\tau > t$

$$\left[e^{gt} + e^{g\tau} - 1\right]^{-\mu/g} = e^{-\mu\tau}\left[1 + e^{-g\tau}(e^{gt} - 1)\right]^{-\mu/g}$$

$$\simeq e^{-\mu\tau}[1 - (\mu/g)e^{-g\tau}(e^{gt} - 1)] \qquad (4.23)$$

Similarly, $\left[e^{gt} - e^{g\tau} + 1\right]^{-\mu/g}$ can be approximated as

$$\left[e^{gt} - e^{g\tau} + 1\right]^{-\mu/g} = e^{-\mu t}\left[1 - e^{-gt}(e^{g\tau} - 1)\right]^{-\mu/g}$$

$$\simeq e^{-\mu t}[1 + (\mu/g)e^{-gt}(e^{g\tau} - 1)] \qquad (4.24)$$

Substituting Eqs. (4.19) to (4.21) into Eq. (4.2), we get

$$(1 - \rho)T(t) = t + \frac{\rho/\mu}{1 - \rho} - \rho\int_0^t T'(\tau)[e^{-\mu t} - (\mu/g)e^{-(g+\mu)t}(e^{g\tau} - 1)]d\tau$$

$$- \rho\int_t^\infty T'(\tau)[e^{-\mu\tau} - (\mu/g)e^{-(g+\mu)\tau}(e^{gt} - 1)]d\tau$$

$$- \rho\int^t T'(\tau)d\tau[e^{-\mu t} + (\mu/g)e^{-(g+\mu)t}(e^{g\tau} - 1)]d\tau$$

$$(4.25)$$

Simplifying Eq. (4.25), we get

$$T(t)(1 - \rho + 2\rho e^{-\mu t}) = t + \frac{\rho/\mu}{1 - \rho} - \rho\int_t^\infty T'(\tau)e^{-\mu\tau}d\tau$$

$$+ p(e^{gt} - 1)\int_t^\infty (\mu/g)e^{-(\mu+g)\tau}T'(\tau)d\tau \qquad (4.26)$$

Differentiating Eq. (4.26) with respect to t yields

$$T'(t)[1 - \rho + \rho e^{-\mu t} + \lambda/g(e^{gt} - 1)e^{-(\mu+g)t}] - 2\lambda T(t)e^{-\lambda t}$$

$$= 1 + \lambda e^{gt}\int_t^{\infty} T'(\tau)e^{-(\mu+g)\tau}d\tau \qquad (4.27)$$

From Eq. (4.27), by setting $t = 0$, and recognizing that $T(0) = 0$, we get the initial condition

$$T'(0) = 1 + \int_0^{\infty} \lambda T'(\tau)e^{-(\mu+g)\tau}d\tau \qquad (4.28)$$

Multiplying e^{-gt} to both sides of Eq. (4.27) and then differentiating it with respect to t, and rearranging the result, we get

$$T'(t)[1 - \rho + \rho e^{-\mu t} + \lambda/g(e^{gt} \cdot 1)e^{-(\mu+g)t}]$$

$$+ T'(t)[-g + g\rho - g\rho e^{-\mu t} - \lambda e^{-\mu t} - e^{-(g+\mu)t}(e^{gt} - 1)\lambda(2 + \mu/g)]$$

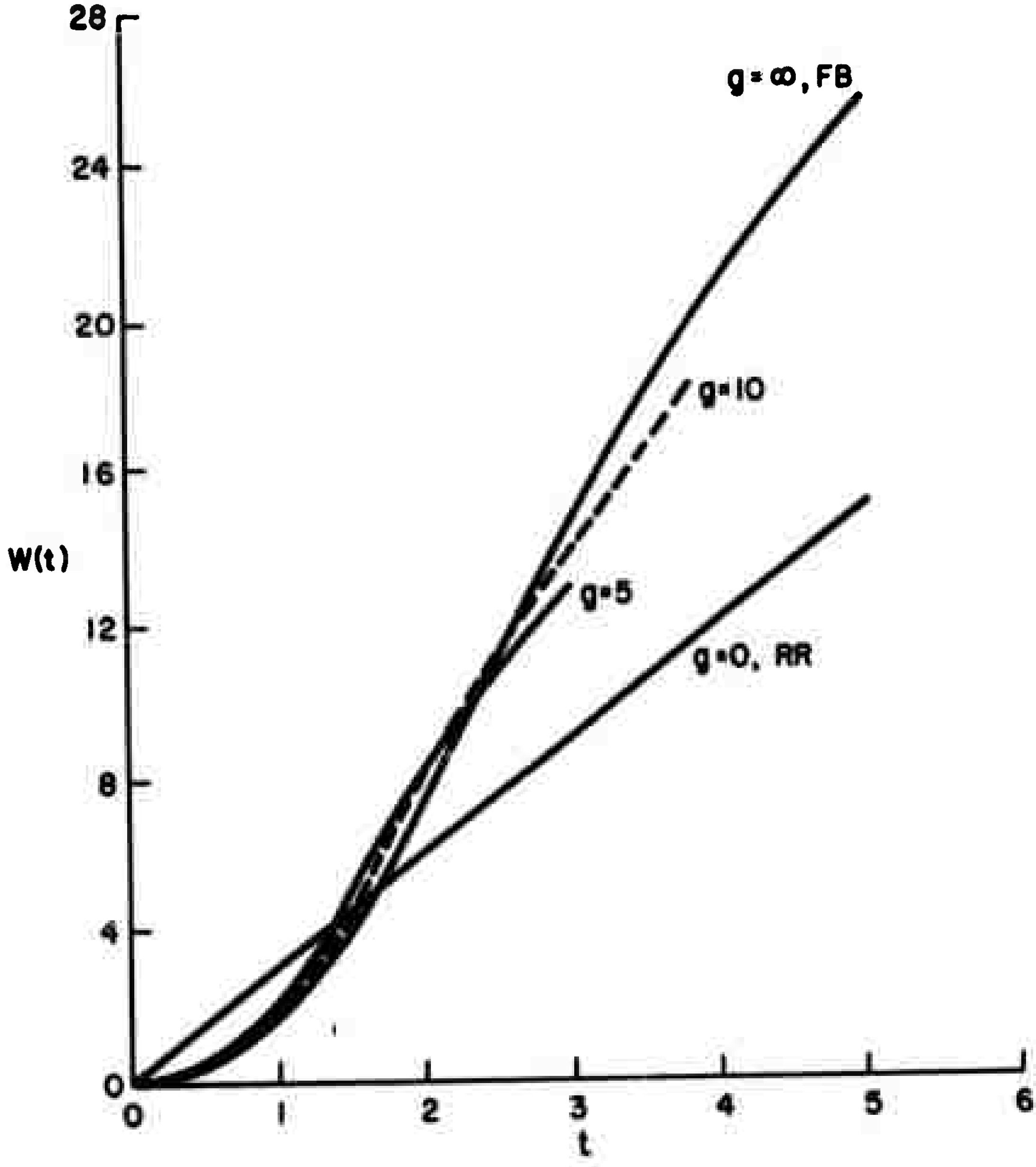
$$+ 2\lambda(\mu + g)e^{-\mu t}T(t) + g = 0 \qquad (4.29)$$

with the initial conditions

$$T(0) = 0$$

$$T'(0) = 1 + \lambda\int_0^{\infty} T'(\tau)e^{-(\mu+g)\tau}d\tau \qquad (4.30)$$

Milne's mechod [37] is used to solve Eq. (4.30). Since the initial condition $T'(0)$ is given in terms of some unknown results, many iterations are needed in order for the results to converge. $W(t)$ is plotted against t on Figure 4-2 with $\lambda = 0.75$ and $\mu = 1$.



WAITING TIME FUNCTIONS PLOTTED AGAINST t FOR DIFFERENT VALUES OF g WITH λ = 0.75 AND μ = 1

FIGURE 4-2

CHAPTER 5

THE AVERAGE NUMBER OF CUSTOMERS IN THE SYSTEM

5.1 Introduction

In Section 2.1.3 we discussed the memoryless property of Markovian process. In particular, the average number of customers in an M/M/1 system was given by

$$\bar{n} = \sum_{n=0}^{\infty} np_n = \frac{\rho}{1 - \rho} \tag{5.1}$$

where p_n is the equilibrium probability of having n customers in the system and is given as

$$p_n = (1 - \rho)\rho^n \qquad n = 0,1,2,3,... \tag{5.2}$$

Equation (5.1) holds true only when no more information about the system is available besides that the arrival and service processes are Markovian. Otherwise, the average number in the system will change according to the additional information. For example, let us assume that the system is not idle when the average is taken; in other words, there is at least one customer in the system, and ask what is the average number of customers in the system under this condition. Equation (5.2) then becomes

$$p_n = (1 - \rho)\rho^{n-1} \qquad n = 1,2,3,... \tag{5.3}$$

and

$$p_0 = 0 \tag{5.4}$$

The average number of customers in the system becomes

$$\bar{n} = \sum_{n=0}^{\infty} np_n$$

$$= \frac{1}{\rho} \sum_{n=0}^{\infty} n(1 - \rho)\rho^n$$

$$= \frac{1}{\rho} \cdot \frac{\rho}{1 - \rho}$$

$$\bar{n} = \frac{1}{1 - \rho} \tag{5.5}$$

which is just the overall average number $\rho/(1 - \rho)$ divided by ρ, the probability that the system is busy.

Both Eqs. (5.1) and (5.5) were obtained under the asumption that no specific information about any single customer was available; all customers were assumed to be identically distributed as far as their interarrival and service times were concerned. In most of the time-sharing models, however, the quantity that is solved for is the average response time in the system conditioned on a tagged customer needing exactly t seconds of service time. In those systems, one of the customers in the system [the tagged customer] behaves differently from others. His service time request becomes deterministic and is no longer drawn from an exponential distribution as are all other service times.

Given this additional information (that one customer will not leave the system before he has obtained exactly t seconds of service), one would expect that the average number of customers in the system will change as a function of this t. In fact it does as is shown by the following theorem in a very simple way.

5.2 The Analytical Results

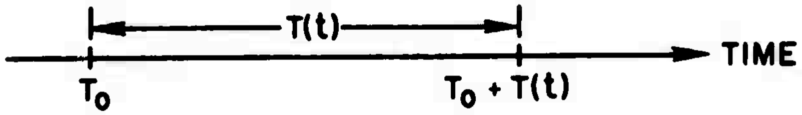
Theorem 5.1

The average number of customers $m(t)$ in an $M/M/1$ system given that one (the tagged) customer has attained t seconds of service is given by

$$m(t) = \frac{1}{1 - \rho} + \lambda T(t) - \mu W(t) \tag{5.6}$$

where $T(t)$ and $W(t)$ are the average response and waiting time, respectively, for the tagged customer.

Proof:



THE REAL TIME AXIS

FIGURE 5-1

The proof of this theorem is very simple. Let us look at the real time axis as shown in Figure 5-1. Let the tagged customer arrive at an arbitrary time instant T_0. Since no extra information about the system is available, there are on the average $1/(1 - \rho)$ customers in the system (including this just arrived tagged customer) at T_0. At time $[T_0 + T(t)]$, the tagged customer has spent $T(t)$ seconds in the system and thus has been in the service facility for an average of t seconds. During the interval $[T_0 \; T_0 + T(t)]$, on the average, there are $\lambda T(t)$ [Poisson arrival] new customers arriving to the system.

There are $\mu[T(t) - t]$ customers leaving the system because the death rate becomes zero when the tagged customer is served (no customer can leave the system when the tagged customer is in the service facility). The average number of the customers $m(t)$ in the system at $T_0 + T(t)$ is then calculated as

$$m(t) = \frac{1}{1 - \rho} + \lambda T(t) - \mu[T(t) - t]$$

$$= \frac{1}{1 - \rho} + \lambda T(t) - \mu W(t) \qquad (5.7)$$

Q.E.D.

Equation (5.7) can be rearranged in one of the following two different forms.

$$m(t) = \frac{1}{1 - \rho} - (\mu - \lambda)W(t) + \lambda t \qquad (5.8)$$

or

$$m(t) = \frac{1}{1 - \rho} - (\mu - \lambda)T(t) + \mu t \qquad (5.9)$$

Since the average number of customers in a busy system with no constraints is $1/(1 - \rho)$, the weighted average of $m(t)$ for all possible t must be equal to $1/(1 - \rho)$. This can be easily shown by using the Conservation Law (Eq. (2.78))

$$\int_0^\infty m(t) dB(t) = \int_0^\infty [\frac{1}{1 - \rho} + \lambda T(t) - \mu W(t)] \mu e^{-\mu t} dt$$

$$= \frac{1}{1 - \rho} + \lambda \int_0^\infty T(t) \mu e^{-\mu t} dt - \mu \int_0^\infty W(t) \mu e^{-\mu t} dt$$

$$= \frac{1}{1 - \rho} + \lambda \frac{1/\mu}{1 - \rho} - \mu \cdot \frac{\rho/\mu}{1 - \rho}$$

$$= \frac{1}{1 - \rho} \qquad (5.10)$$

For the RR system, the average response time is linearly proportional to t and is given by

$$T_{RR}(t) = \frac{t}{1 - \rho} \qquad (5.11)$$

substituting Eq. (5.11) into Eq. (5.9), the average number of customers in a RR system given that a customer in the system has obtained t seconds of service time becomes

$$\begin{aligned} m_{RR}(t) &= \frac{1}{1 - \rho} - (\mu - \lambda)\frac{t}{1 - \rho} + \mu t \\ &= \frac{1}{1 - \rho} - \mu t + \mu t \\ &= \frac{1}{1 - \rho} \qquad (5.12) \end{aligned}$$

which is independent of t and always assumes the same value. As a consequence of Eq. (5.12), we get the following theorem:

Theorem 5.2

For any scheduling algorithm and its corresponding average response time function $T(t)$, if $T(t)$ assumes the same value as the average response time function $T_{RR}(t)$ (for the RR algorithm) at some $t = t_i$, then at this point t_i (that is, given the information that one of the customers in either system has attained exact t_i seconds of service time), the average number of customers in both systems equal to $1/(1 - \rho)$, i.e. if

$$T(t_i) = T_{RR}(t_i)$$

then

$$m(t_i) = \frac{1}{1 - \rho}$$

Proof:

If we substitute

$$T_{RR}(t) = \frac{t}{1 - \rho}$$

into Eq. (5.7), then $m_{RR}(t)$ becomes

$$m_{RR}(t) = \frac{1}{1 - \rho} \tag{5.13}$$

the average number of customers in a constant in the RR system, independent of t.

Now, if for some scheduling algorithm which gives an average response time $T(t)$, it intersects with $T_{RR}(t)$ at $t = t_i$, then

$$T(t_i) = T_{RR}(t_i) = \frac{t_i}{1 - \rho}$$

and

$$W(t_i) = W_{RR}(t_i) = \frac{\rho t_i}{1 - \rho} \tag{5.14}$$

Substituting Eq. (5.14) into Eq. (5.7), we get

$$m(t_i) = \frac{1}{1 - \rho} + \frac{\lambda t_i}{1 - \rho} - \frac{\rho t_i}{1 - \rho} \cdot \mu$$

$$= \frac{1}{1 - \rho} \tag{5.15}$$

Q.E.D.

5.3 Examples

Equation (5.7) is good for all M/M/1 systems, be it infinite-quantum or processor-sharing. In this section we want to plot the average number of customers for different scheduling algorithms to demonstrate the nature of Eq. (5.7).

5.3.1 First-Come-First-Served (FCFS) System

The average waiting time is given by Eq. (2.19) as

$$W(t) = \frac{\rho/\mu}{1 - \rho} \tag{5.16}$$

substituting Eq. (5.16) into Eq. (5.8) gives

$$\begin{aligned} m(t) &= \frac{1}{1 - \rho} - (\mu - \lambda)\frac{\rho/\mu}{1 - \rho} + \lambda t \\ &= \frac{1}{1 - \rho} - \rho + \lambda t \\ &= \frac{1 - \rho + \rho^2}{1 - \rho} + \lambda t \end{aligned} \tag{5.17}$$

m(t) is plotted against t in Figure 5-2 with $\lambda = 0.75$ and $\mu = 1$. The average number of customers increases with t because once the tagged customer enters the service facility, he will occupy the facility all by himself and thus block all those who come to the system later than he does. The number of "blocked" customers increases with rate λ, and if t goes to infinity, so goes the average number of customers in the system.

5.3.2 Round Robin (RR) System

As proved by Theorem 5.2, the average number m(t) is always equal to $1/(1 - \rho)$ in the RR system. It serves as a natural reference algorithm for all other algorithms.

5.3.3 Selfish Round Robin (SRR) System

The average response time for the SRR system is given by Eq. (2.35) as

$$T(t) = \frac{1/\mu}{1 - \rho} + \frac{(t - 1/\mu)}{1 - \rho(1 - \beta/\alpha)} \qquad (5.18)$$

Substituting Eq. (5.18) into Eq. (5.9), m(t) is obtained as

$$m(t) = \frac{1}{1 - \rho} - (\mu - \lambda)\left[\frac{1/\mu}{1 - \rho} + \frac{(t - 1/\mu)}{1 - \rho(1 - \beta/\alpha)}\right] + \mu t$$

$$= \frac{\rho}{1 - \rho} + \frac{1 - \rho}{1 - \rho(1 - \beta/\alpha)} + \mu t \frac{\rho(\beta/\alpha)}{1 - \rho(1 - \beta/\alpha)} \qquad (5.19)$$

m(t) is plotted ageinst t for the case of $\beta/\alpha = 0.2$ in Figure 5-2 with $\lambda = 0.75$ and $\mu = 1$. The behavior of the SRR system is again bounded by those of the RR and the FCFS systems. The average number goes to infinity but at a slower rate than that for the FCFS system because the "blocking" effect is not as severe in the SRR system. All SRR curves give $m(t) = 1/(1 - \rho)$ for $t = 1/\mu$.

5.3.4 Foreground Background (FB) System

Substituting Eg. (2.50) into Eg. (5.9), m(t) for FB system is expressed as

$$m(t) = \frac{1}{1 - \rho} - (\mu - \lambda)\left[\frac{(\rho/\mu)(1 - e^{-\mu t} - \mu t e^{-\mu t})}{(1 - \rho + \rho e^{-\mu t})^2} + \frac{t}{1 - \rho + \rho e^{-\mu t}}\right] + \mu t$$

$$(5.20)$$

this is plotted against t in Figure 5-2 with $\lambda = 0.75$ and $\mu = 1$. When t is small, the tagged customer receives service at a rate larger than most of the other customers in the system, thus causing a smaller death rate and the number of customers in the system increases. As his attained service time accumulates, the tagged customer constantly loses priority and gets served at decreasing rate. The death rate of the system increases as t increases, and there are fewer and fewer customers remaining in the system. When t gets very large, the only time the tagged customer can get any service is when he is the only customer in the system, thus the average number goes down to 1 as t goes to infinity.

5.3.5 Multilevel Processor-Sharing Models

Let us use a two-level model for our example. FCFS algorithm is used for both levels, and let x denote the break point. Combining Eqs. (2.57) and (5.7), m(t) is given by

$$m(t) = \begin{cases} \dfrac{1}{1-\rho} - (\mu - \lambda)[t + \dfrac{(\rho/\mu)(1 - e^{-\mu x} - \mu x e^{-\mu x})}{1 - \rho + \rho e^{-\mu x}}] + \mu t & t \leq x \\ \\ \dfrac{1}{1-\rho} - (\mu - \lambda)[\dfrac{\rho/(1-\rho) + t}{1 - \rho + \rho e^{-\mu x}} + \mu t & t \geq x \end{cases} \qquad (5.21)$$

In Figure 5-2, m(t) for x = 2 and x = 3 are plotted against t with $\lambda = 0.75$ and $\mu = 1$. m(t) increases at the same rate as in a FCFS system when t is smaller than x because the tagged customer is the only one in the service facility and he blocks all the late comers as well as all those in the system who have thus far received more than x seconds of service. Having attained x seconds of service, the tagged customer moves to the lower priority group, and he must wait there until the system serves all those customers previously blocked by him up to x seconds each as well as those who have received more than x seconds of service time and were preempted by him when he entered the service in the lower level. This accounts for the large drop of customers at t = x . Even after he regains control of the service facility at some later time, the tagged customer is always subjected to preemption whenever a new customer arrives. The tagged customer still can block those customers who need more than x seconds of service after they have gained that much, but the rate of increase of customers in the system is rather low if x is considerably larger than the average service request $1/\mu$.

5.3.6 Tight Upper and Lower Bounds

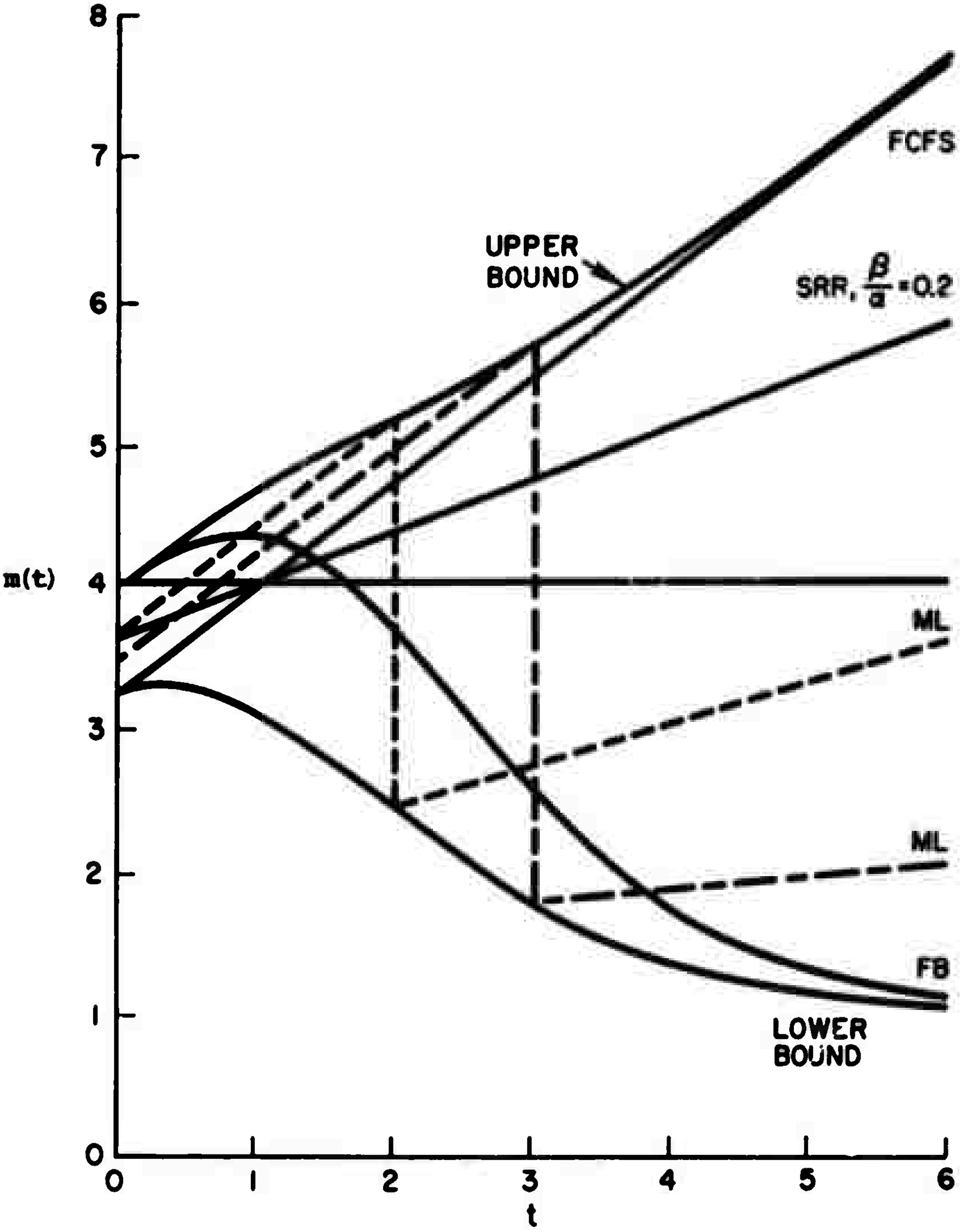
In Chapter 7, tight upper and lower bounds as measured by the response time function are derived for the processor-sharing models for time-shared systems. They are given by, respectively,

$$T_u(t) = \frac{\rho/\mu}{(1 - \rho + \rho e^{-\mu t})(1 - \rho)} + \frac{t}{1 - \rho + \rho e^{-\mu t}} \qquad (5.22)$$

$$T_\ell(t) = \frac{\rho/\mu(1 - e^{-\mu t} - \mu t e^{-\mu t})}{1 - \rho + \rho e^{-\mu t}} \qquad (5.23)$$

In Eq. (5.9), $T(t)$ appears only once and has a minus sign in front of it. For a given t, if we substitute the tight upper bound $T_u(t)$ of $T(t)$ into Eq. (5.9), we get the tight lower bound of the number of customers in the system. Similarly, the tight upper bound of $m(t)$ is obtained by substituting the tight lower bound $T_\ell(t)$ of the response time functions into Eq. (5.9). The tight bounds of the average number of customers in the system are plotted against t in Figure 5-2 with $\lambda = 0.75$ and $\mu = 1.0$. As is shown by Figure 5-2, as t goes to infinity the lower bound goes to 1; and the upper bound approaches the average number for the FCFS system asympotically.

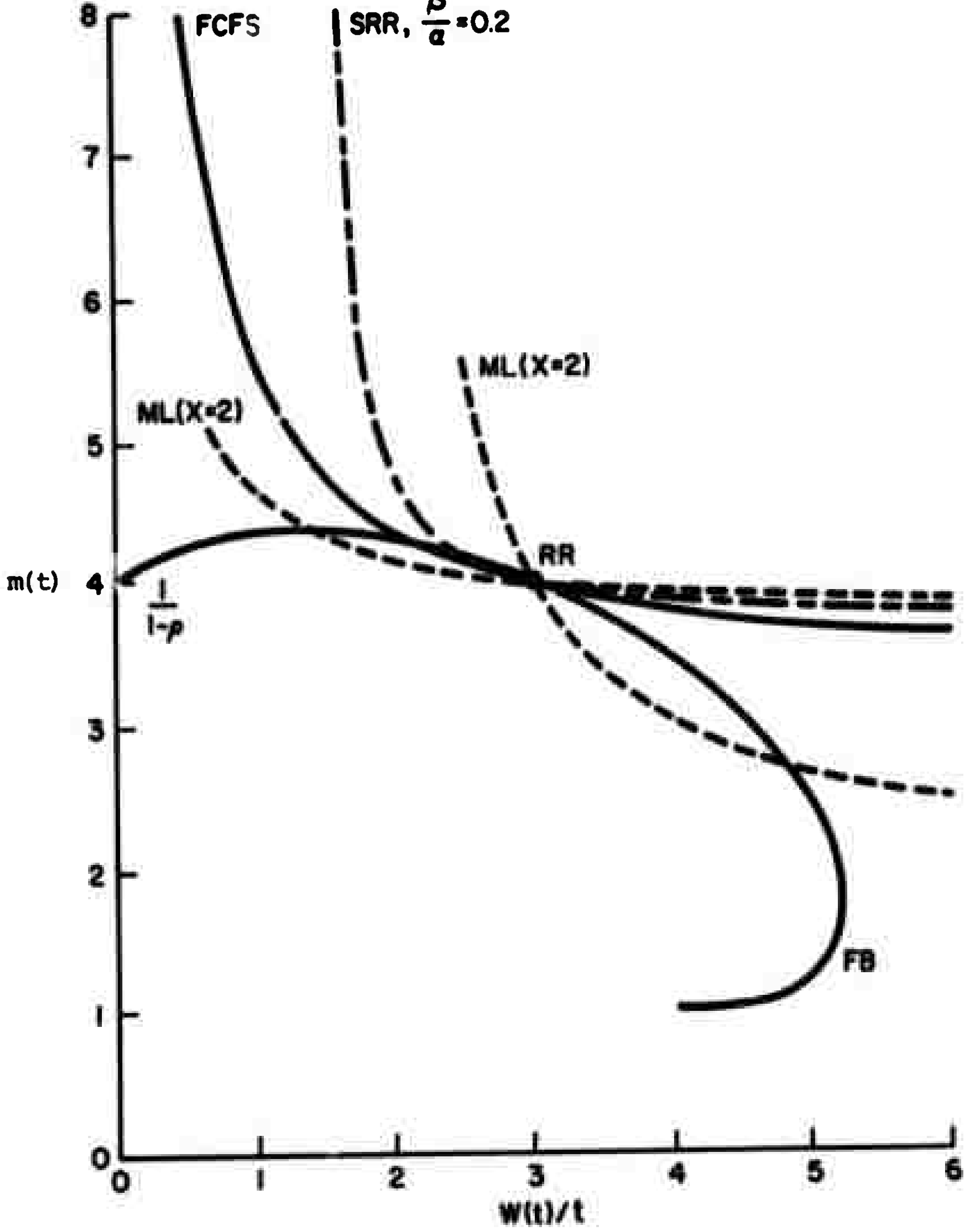
In Figure 5-3, $m(t)$ is plotted against $W(t)/t$ which denotes how large a price (in terms of wasted time) a customer must pay in order to get a unit of service with $\lambda = 0.75$ and $\mu = 1.0$ for different



TIME-VARYING AVERAGE NUMBER OF CUSTOMERS PLOTTED AGAINST t FOR DIFFERENT ALGORITHMS WITH $\lambda = 0.75$, $\mu = 1$

FIGURE 5-2

algorithms. For RR, both $m(t)$ and $W(t)/t$ are constants, it says that every customer pays the same amount for a unit of servi.e time indeppendent of his attained service, and the average number of customers in the system does not change with t or any function of t . For all other algorithms, $m(t)$ is larger than $1/(1 - \rho)$ when $W(t)/t$ is smaller than $\rho/(1 - \rho)$. The interpretation of this phenomenon is that with $W(t)/t < \rho/(1 - \rho)$, the tagged customer has been treated better than the average (as represented by the RR system) so that his occupancy in the service facility has been somewhat longer than that he would encounter in a RR system. As a consequence of this, the average death rate (rate of departures) of the customers in the system is somewhat lower than the average arrival rate during his stay in the system (since for RR system, these two rates are the same), thus the average number of customers in the system goes up. On the other hand, $m(t)$ decreases as $W(t)/t$ goes above the constant $\rho/(1 - \rho)$ since the average departure rate from the system would then be higher than the average arrival rate to the system. For the FB system (refer to Figure 3.9 where $W(t)$ is plotted against t), the function $W(t)/t$ increases with t until it hits its maximum value (at $t = 5.25$ for $\lambda = 0.75$ and $\mu = 1$). It then starts to decrease as t increases and approaches $1/(1 - \rho)$ as t goes to infinity; at the same time the average number of customers in the system goes down to 1 asympotically. This accounts for the hook-shaped curve for the FB system in Figure 5-3.



TIME-VARYING AVERAGE NUMBER OF CUSTOMERS PLOTTED AGAINST W(t)/t FOR DIFFERENT ALGORITHMS WITH $\lambda = 0.75$, $\mu = 1.0$

FIGURE 5-3

CHAPTER 6

TIGHT BOUNDS ON THE AVERAGE RESPONSE TIME*

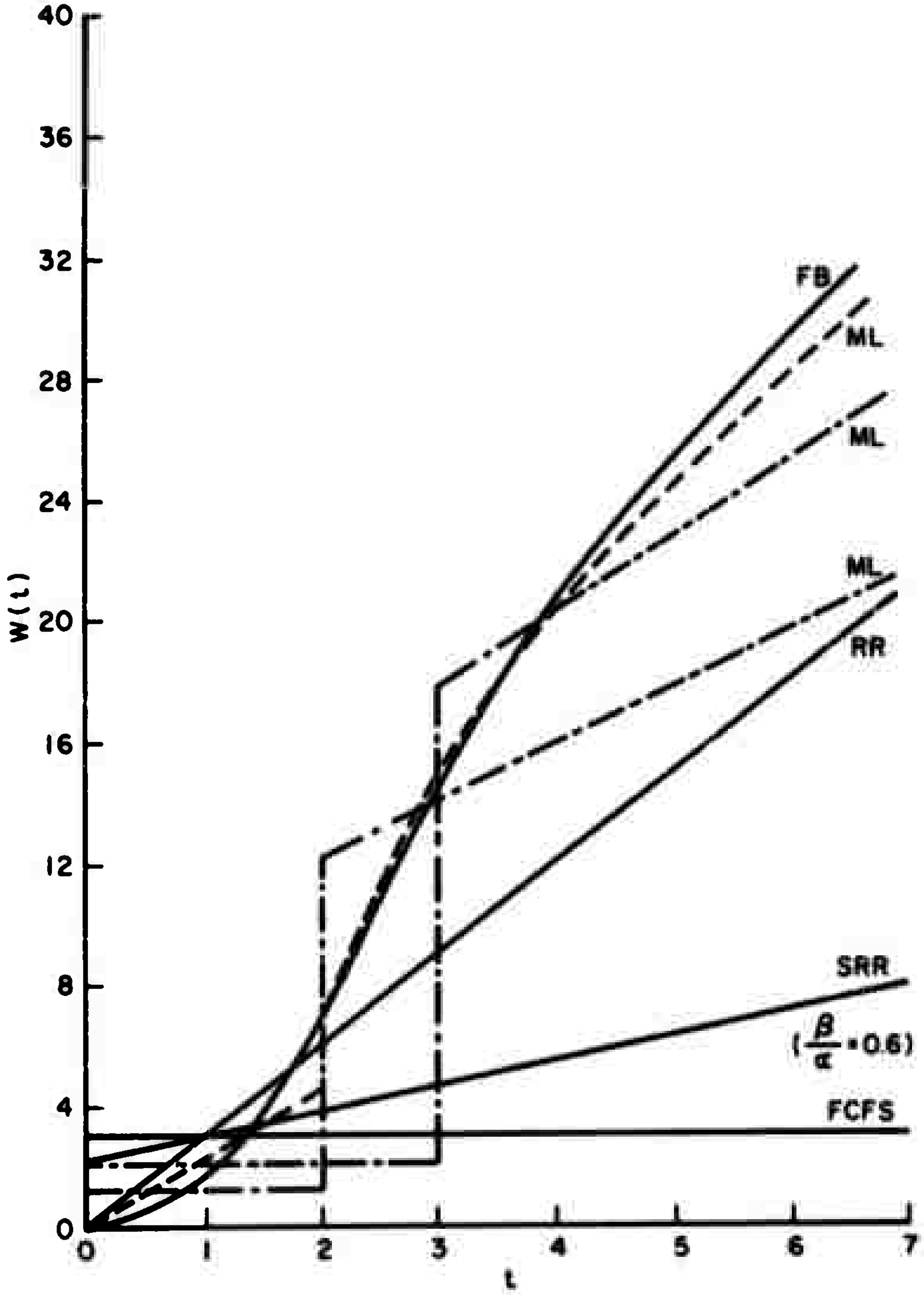
6.1 Introduction

In the previous chapters of this dissertation we discussed a few models of time-shared computer systems. By slightly changing the set of assumptions for those systems, more models could be constructed and more analytical results could be obtained. As a result of this flood of results, it is natural that we should seek some order. For example, do there exist any invariants in behavior? Can we bound the possible range of performance regardless of structure? What constitute feasible solutions for these systems? These, and many more, are reasonable inquiries to make amidst the confusion of results.

In this chapter we try to answer some of the questions. Our focus is on a class of processor-sharing models of time-shared computer systems. For these processor-shared systems, it is useful to display, in one figure, the wasted time $W(t)$. This we do in Figure 6-1 for the case of expontial service with $\lambda = 0.75$ and $\bar{t} = 1.0$ (thus $\rho = 0.75$). We purposely superimpose the performance curves for many scheduling disciplines. We are confronted with quite a selection of possible performance functions! For these systems we are able to state a monotonicity property, a conservation law, and tight upper and lower bounds on the system performance as measured by average response time.

*These results were obtained in collaboration with L. Kleinrock and R. R. Muntz [44].

Preceding page blank



A SET OF RESPONSE CURVES FOR M/M/1, μ = 1.0, λ = 0.75

FIGURE 6-1

It is worthwhile mentioning that numerous papers have recently been published which address themselves to bounds, inequalities and approximate solutions to general queueing systems. Among these are Marshall [38,39], Kingman [40], Iglehart [41], Daley and Moran [42], and Gaver [43], to mention a few.

6.2 The Analytical Results

In this section we present results concerning the response functions (W(t)) which are feasible when the scheduling discipline is based only on attained service time and elapsed waiting time of jobs. In Theorem 6.1 we state a monotonicity property for $W(t)$. In Theorem 6.2 we give a conservation relationship which the response function must satisfy. In Theorem 6.3 and 6.4 tight lower and upper bounds are derived. As a result of Theorem 6.4, a necessary condition for $W'(t)$ is obtained in Theorem 6.5.

Theorem 6.1 $W(t)$ is a nondecreasing function of t or equivalently

$$W'(t) \equiv \frac{dW(t)}{dt} \geq 0 \qquad (6.1)$$

Proof: See Appendix C

Theorem 6.2 There is a conservation law that $W(t)$ has to satisfy, namely

$$\int_0^\infty W(t)[1 - B(t)]dt = \frac{\rho\overline{t^2}}{2(1-\rho)} \qquad (6.2)$$

For $T(t)$, the conservation law becomes

$$\int_0^\infty T(t)[1 - B(t]dt = \frac{\overline{t^2}}{2(1-\rho)} \qquad (6.3)$$

Proof: See Appendix C

We refer to Eqs. (6.2) and (6.3) as Conservation Laws since they are based on the conservation of average unfinished work in the system. This places an integral constraint on $W(t)$ (and $T(t)$) as a second necessary condition, regardless of scheduling algorithm. The implication of the conservation law may be seen by recognizing that $[1 - B(t)]$ is a non-increasing function of t. Thus, if one had a given $W(t)$ as a result of some scheduling algorithms, and then changed the algorithm so as to reduce $W(t)$ over some interval $(0,t_0)$, then the conservation law would require that the new $W(t)$ be considerably above the old value for some range above t_0. This follows since the weighting factor, $1 - B(t)$, is smaller for large t.

With the help of Theorems 6.1 and 6.2, we now proceed to prove the main theme of this chapter.

Theorem 6.3 The lower bound $W_\ell(t)$ of waiting time functions is given by the waiting time for the FCFS discipline with the service time distribution truncated at t, namely

$$W_\ell(t) = \frac{\lambda \, \overline{t^2}_{<t}}{2(1 - \rho_{<t})} \tag{6.4}$$

Note: that $W_\ell(0) = 0$ and that $W_\ell(\infty) = W_{FCFS}$ (the average waiting time for the FCFS system as given in Section 2.2.1); also $W'_\ell(0) = W'_\ell(\infty) = 0$.

Proof: See Appendix C

Theorem 6.4 The upper bound $W_u(t)$ of waiting time functions is given as

$$W_u(t) = \frac{\lambda \, \overline{t^2}}{2(1 - \rho_{<t})\ (1 - \rho)} + \frac{t \cdot \rho_{<t}}{1 - \rho_{<t}} \tag{6.5}$$

One scheduling discipline which gives $W_u(t)$ is a two-level system with both levels served FCFS and switching point (see Chapter 6) at t. Note that $W'_u(0) = W_\ell(\infty) = W_{FCFS}$, that $W'_u(0) = 0$ and that $W'_u(\infty) = \frac{\rho}{1 - \rho}$

Proof: See Appendix C

As a consequence of Theorem 6.2 and Theorem 6.4, we get the following necessary condition for $W'(t)$.

Theorem 6.5 For a response time function $W(t)$ which is continuously differentiable, $W'(t) = \frac{dW(t)}{dt}$ can not be monotonically non-decreasing with t.

Proof: See Appendix C

6.3 Examples

Four examples are given in this section to demonstrate the nature of the tight bounds we have obtained. As a performance measure, the equilibrium average waiting times, $W(t)$, are plotted as a function of t. We begin with the M/M/1 system (i.e., Poisson arrivals and exponential service). The response functions of Figure 6-1 are given again in Figure 6-2 with the upper and lower bounds superimposed. At $t = 0$, the upper bound and FCFS start at the same point because, under the constraint of the conservation law, no other scheduling algorithm can give longer average waiting time at $t = 0$ than FCFS. The upper bound approaches the FB response asymptotically as t approaches infinity. Therefore, a customer with a very long requested service time (as compared to the mean) cannot be delayed much more than he is with FB. The lower bound starts at zero (as does the FB curve), increasing less rapidly with t than the upper bound. It approaches the FCFS curve asymptotically as t goes to infinity. Thus we note that the least discriminating scheduling algorithm (FCFS) touches the upper bound at $t = 0$ and forms the asymptote for the lower bound as t approaches infinity; conversely, the most discriminating scheduling algorithm (FB) touches the lower bound at $t = 0$ and forms the asymptote for the upper bound as t approaches infinity. The above-mentioned behavior of the upper and lower bounds applies not only for the M/M/1 system, but also holds true for any M/G/1 system in general, although the rate of convergence for the bounds to their respective limits varies for different service distributions.

For the second example we choose the system $M/E_2/1$. In this system we have

$$\frac{dB(x)}{dx} = (2\mu)^2 xe^{-2\mu x} \qquad x \geq 0 \qquad (6.6)$$

with mean service time equal to $1/\mu$; the second moment of this distribution is $3/2\mu^2$. Because the second moment is smaller than that of the exponential distribution (whose value is $2/\mu^2$), the bounds are tighter in this example than the M/M/1 case, just as one would expect. Figure 6.3 shows the behavior of this system with $\mu = 1$ and $\lambda = 0.75$. It is obvious from the figure that for $t > 5/\mu$, the upper ana lower bounds have essentially reached their asymptotic form.

In the third example we show the bounds for the $M/H_2/1$ system, where H_2 stands for hyperexponential service distribution with

$$\frac{dB(x)}{dx} = 0.5\mu_1 e^{-\mu_1 x} + 0.5\mu_2 e^{-\mu_2 x} \qquad x \geq 0 \qquad (6.7)$$

We chose $\mu_1 = 5\mu$, $\mu_2 = (5/9)\mu$, resulting in a mean service time of $1/\mu$. The second moment of this distribution is $82/25\mu^2$. Figure 6-4 shows the behavior of the $M/H_2/1$ system with $\mu = 1$ and $\lambda = 0.75$. The upper and lower bounds approach to their respective limits at a slower rate than either M/M/1 or $M/E_2/1$ because of the larger second moment.

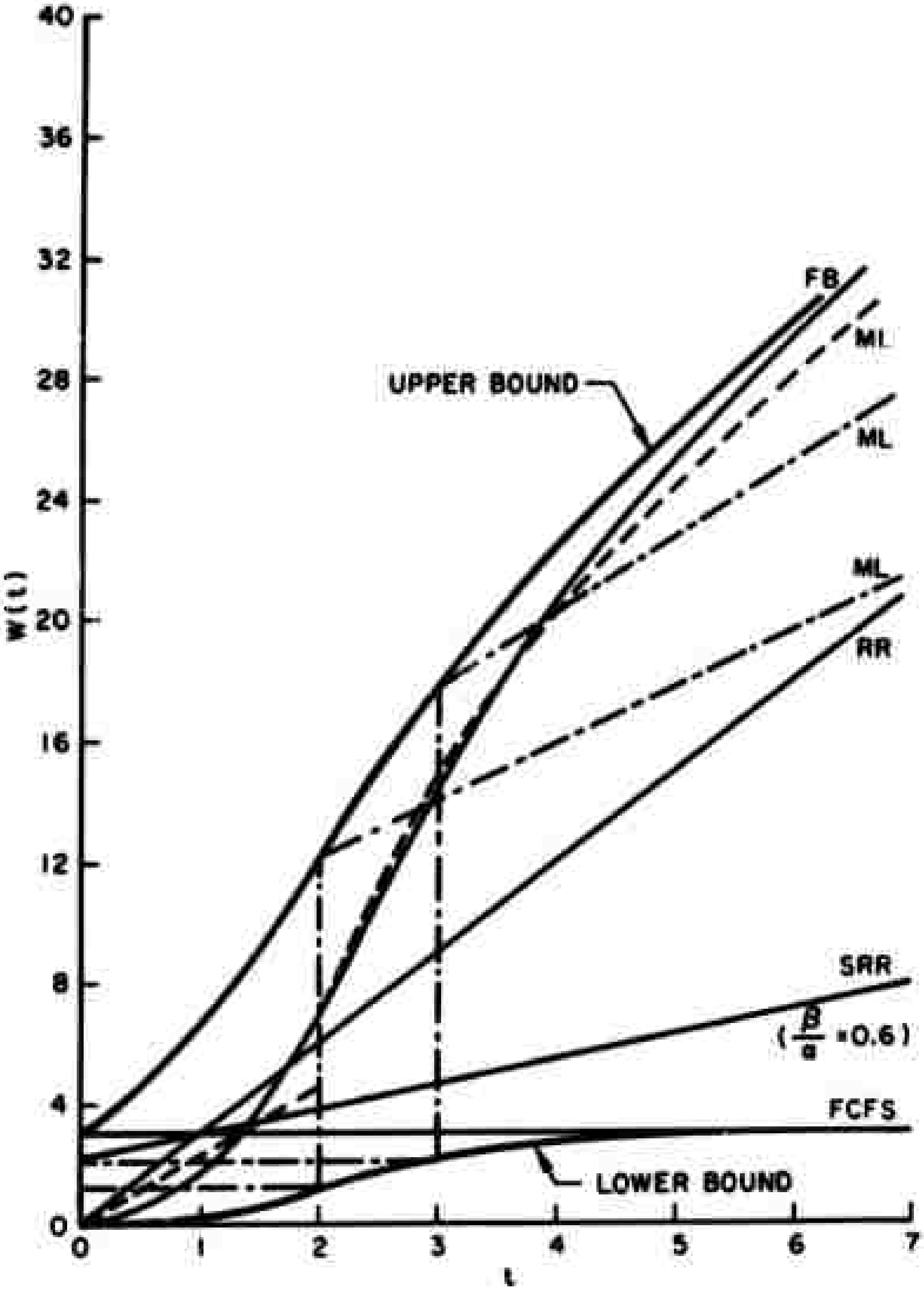
For our last example we choose the system M/u/1 where u stands for uniform service distribution. For this particular example we have

$$\frac{dB(x)}{dx} = \begin{cases} 0.25 & 2 \leq x \leq 6 \\ 0 & \text{otherwise} \end{cases} \qquad (6.8)$$

and $\lambda = 0.1875$, $\bar{t} = 4.0$, $\rho = 0.75$. Figure 6-5 shows the behavior of this system. Notice that when $t \geq 6$, the upper bound coincides exactly with the FB curve and that the lower bound coincides exactly with the FCFS curve. The probability of having any customer requesting more than six seconds of service in this example is, of course, equal to zero.

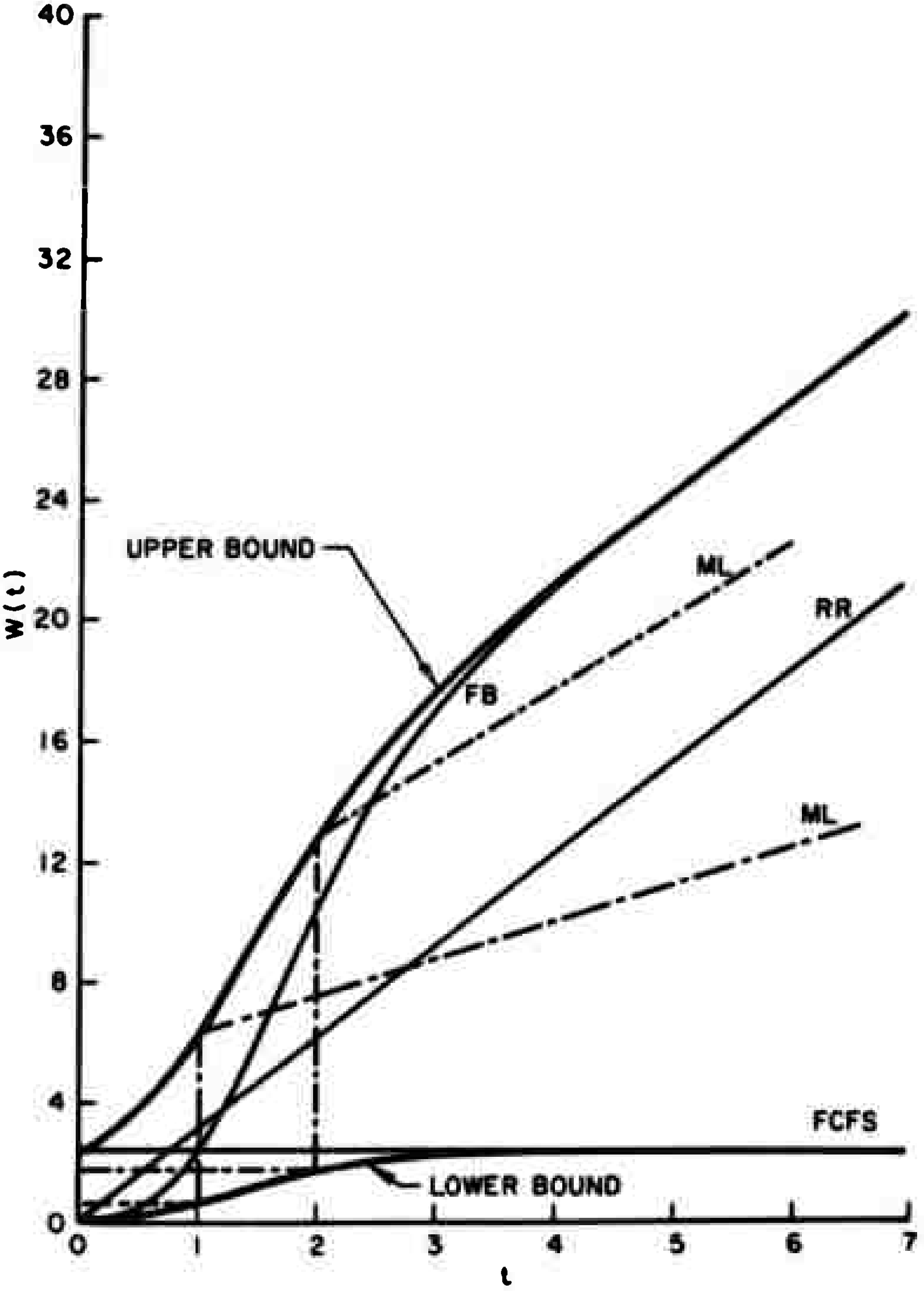
Another performance measure, W(t)/t, is given in Figure 6-6 for theM/M/1 case and is of interest to us, since (as mentioned in Chapter 5) it gives some feeling for how large a price (in terms of wasted time) a customer must pay in order to get a unit of service. For the case of RR, this measure is a constant; thus each customer has the same penalty rate, regardless of his service time. In this sense, everyone is treated equally in the RR system. The curve representing FCFS is monotonically decreasing with t, and so the longer jobs pay at a smaller penalty rate. System users might then attempt to "pool" their requests to take advantage of this "quantity discount." Another extreme example is provided by FB; W(t)/t increases rapidly when t is small, then drops slowly to a constant $(\rho/(1 - \rho))$. A customer with a long request can do better by breaking his job into smaller independent jobs and submitting them separately to the system (if this is possible) because then the average waiting time per unit of service time will be greatly reduced.

Figure 6-7 shows the range of the bounds for the M/M/1 system with $\rho = 0.75$, 0.5 and 0.25, respectively. As can be seen, the region included between the upper and lower bounds for a particular utilization factor ρ depends heavily on ρ; the larger the value of ρ, the greater is the vertical separation between the two bounds, thus allowing larger variation of the mean waiting times for different scheduling algorithms.



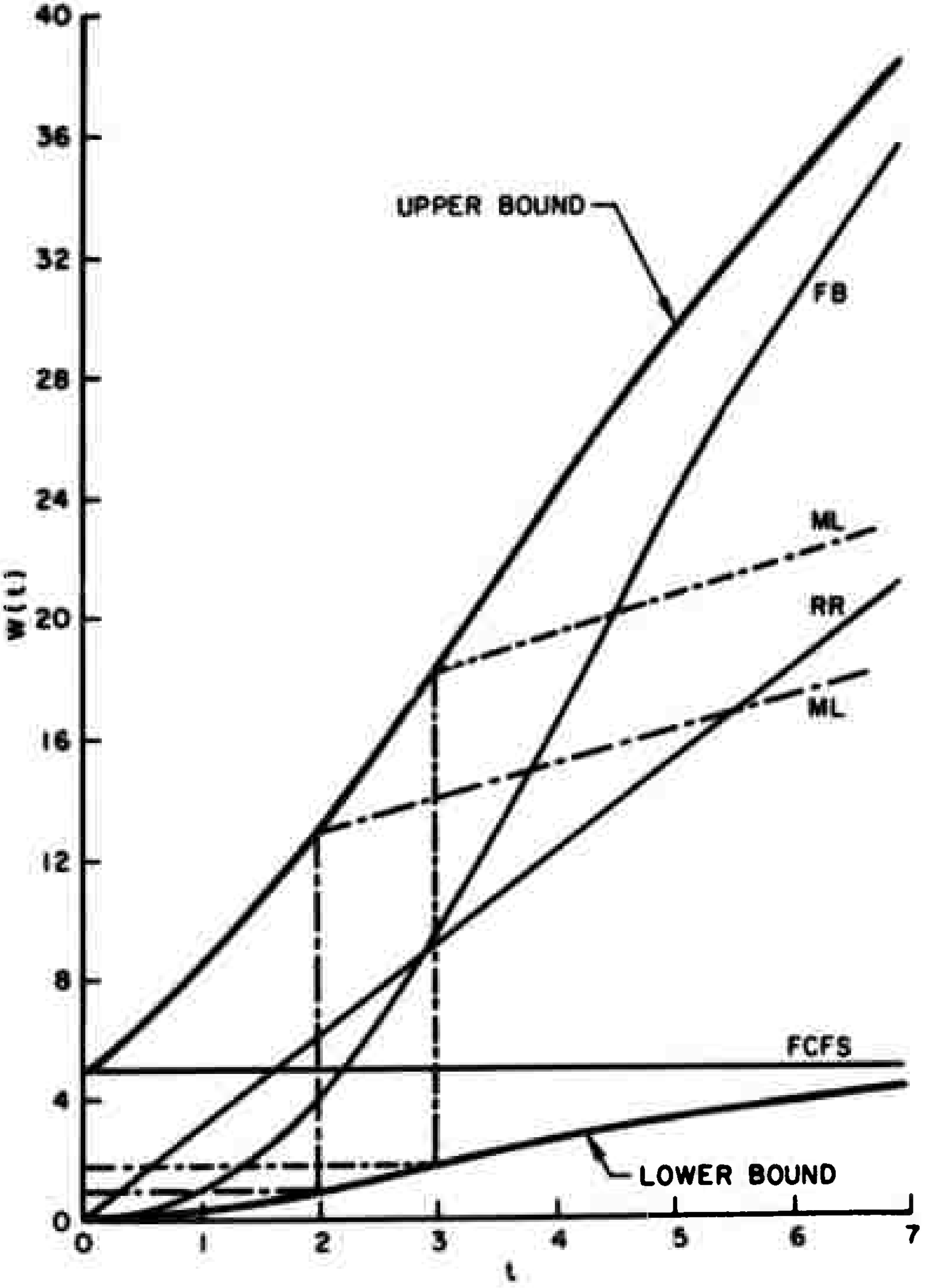
BOUNDS ON RESPONSE FOR M/M/1, $\mu = 1$, $\lambda = 0.75$, $\rho = 0.75$

FIGURE 6-2



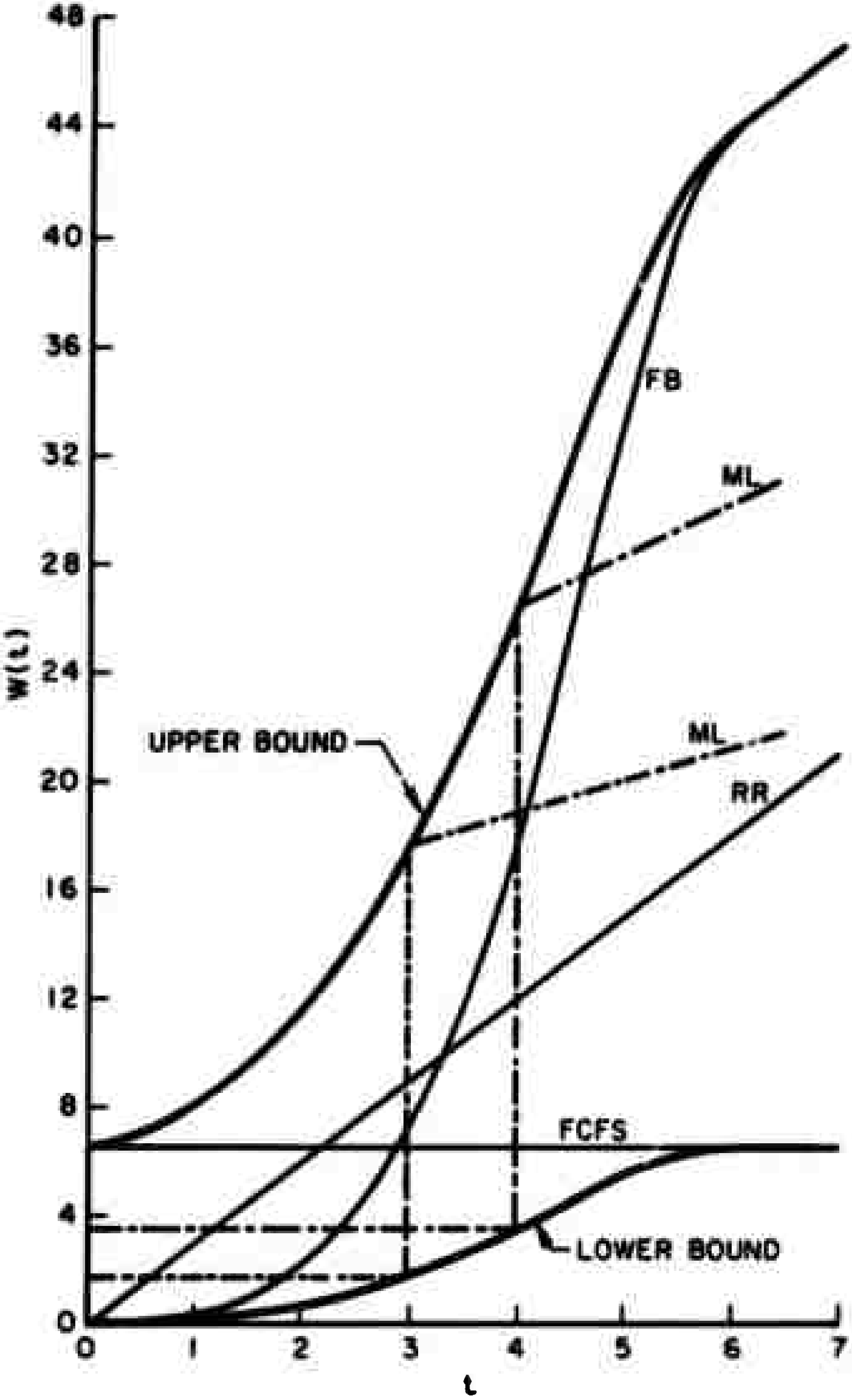
BOUNDS ON RESPONSE FOR $M/E_2/1$, $\mu = 1$, $\lambda = 0.75$, $\rho = 0.75$

FIGURE 6-3



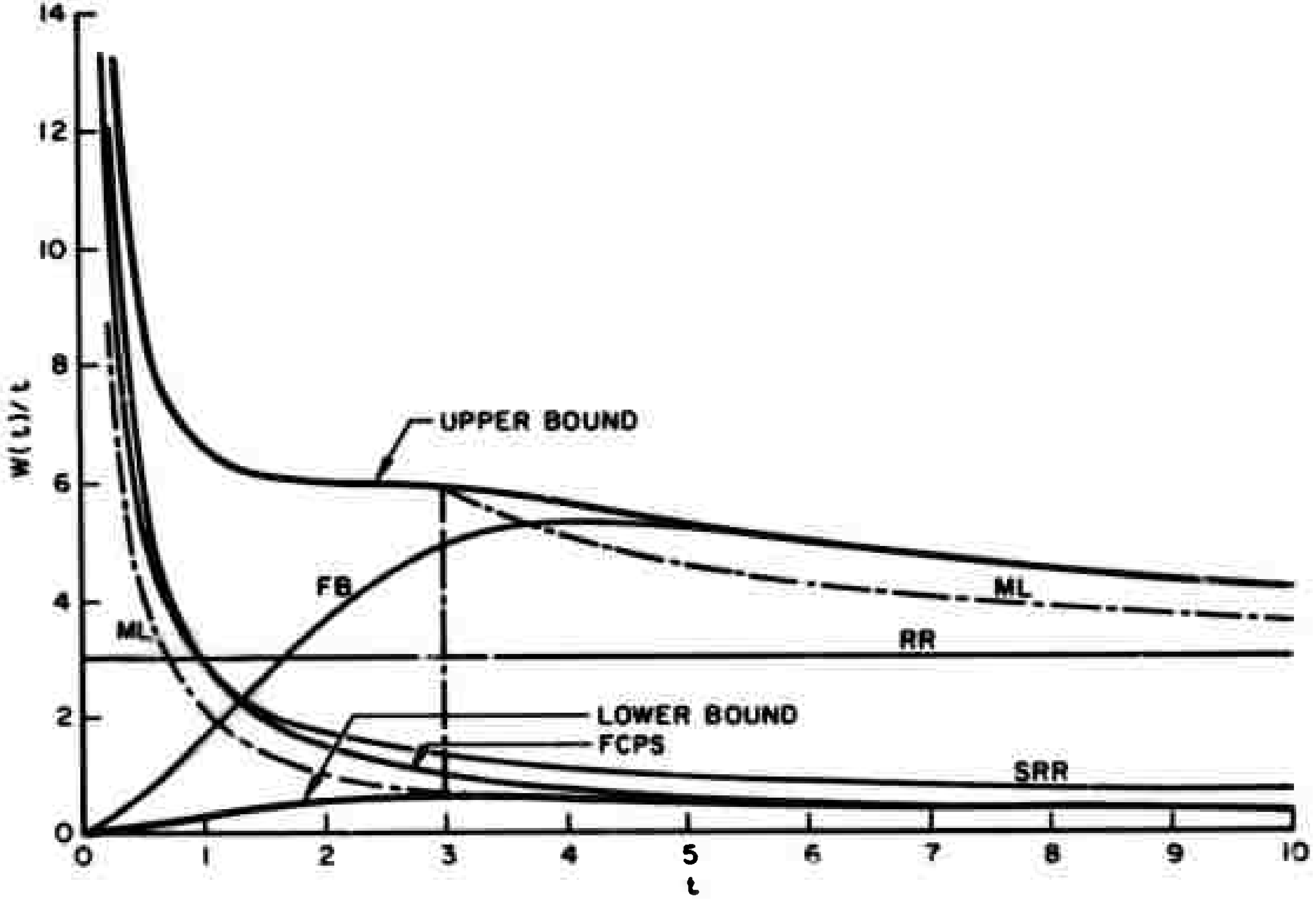
BOUNDS ON RESPONSE FOR $M/H_2/1$, $\mu = 1$, $\lambda = 0.75$, $\rho = 0.75$

FIGURE 6-4



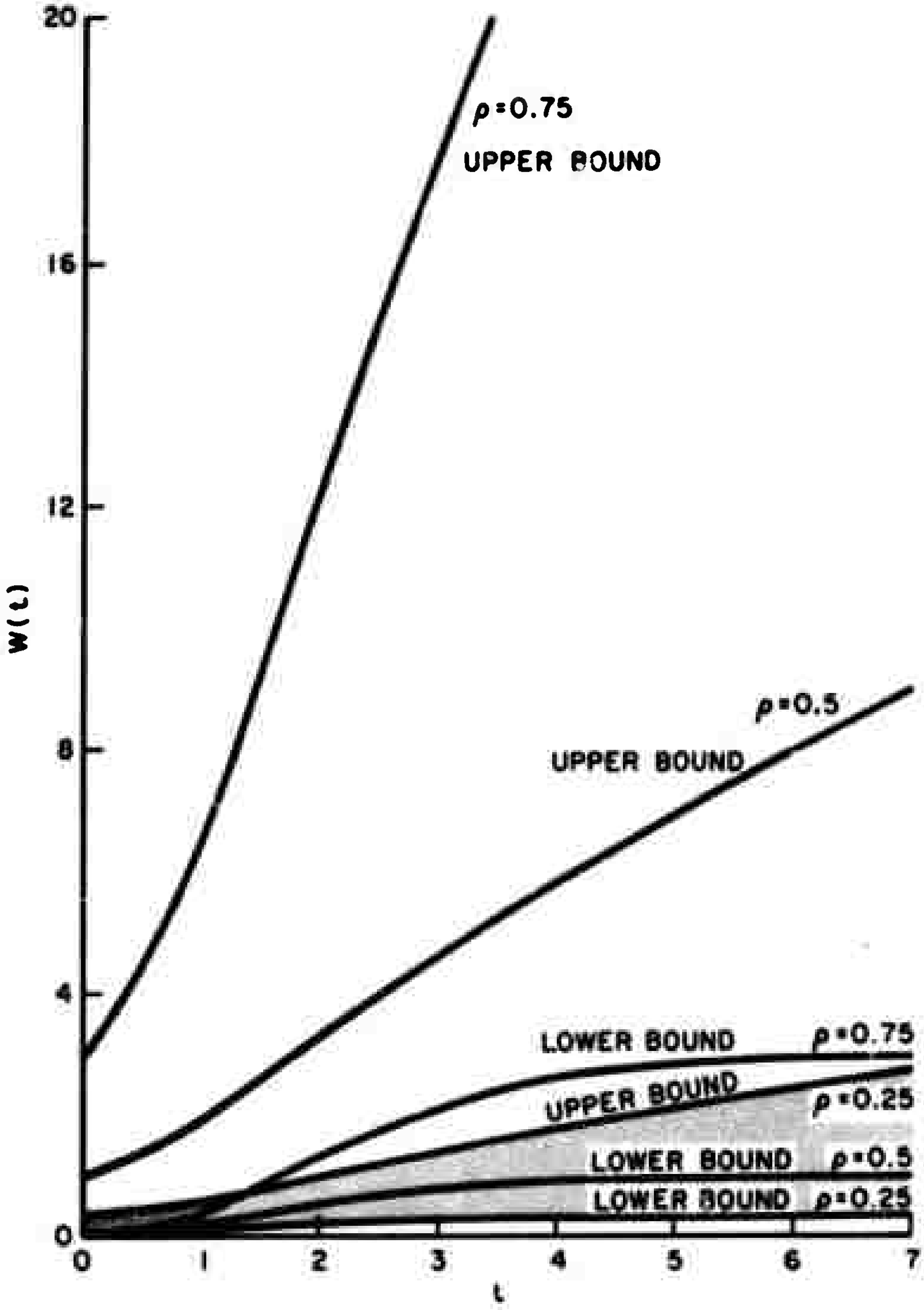
BOUNDS ON RESPONSE FOR M/U/1, $\mu = 0.25$, $\lambda = 0.1875$, $\rho = 0.75$

FIGURE 6-5



BOUNDS ON PENALTY RATE FOR M/M/1, $\nu = 1.0$, $\lambda = 0.75$, $\rho = 0.75$

FIGURE 6-6



VARIATION OF BOUNDS FOR M/M/1 WITH ρ = 0.25, 0.50, AND 0.75

FIGURE 6-7

CHAPTER 7

CONCLUSION AND SUGGESTED AREAS FOR FUTURE RESEARCH

In Section 2.2, we made a survey of some of the results in modelling and analysis of scheduling algorithms that provided for us the starting point of this research. In Chapters 3 and 4, as an extension of this line of work, we modelled and analyzed the family of selfish scheduling algorithms and a new family of algorithms whose perform--ance ranges between that of the RR and the FB systems. The emphasis of these algorithms is to introduce parameters into the models so that various degrees of freedom can be obtained by adjusting these parameters. It is now possible to go from the algorithm (FCFS) which shows no discrimination with regard to job length to that discipline which shows maximum discrimination (FB) on job length among customers in a system. We are able to show models whose performance ranges in between these two extremes on a continuum basis. In Chapters 5 and 6, we answered some of the fundamental questions regarding the existence of order and structure in the analytic results for time-shared computer systems. In particular, conditional average number of customers in the system for different scheduling algorithms is calculated and tight upper and lower bounds are obtained for the class of processor-sharing model queuing systems.

Since we have limited ourselves to the modelling and analysis of only the processor-sharing systems in this dissertation, it is natural then to ask as our first question what kind of extension needs to be done in regard to this area. In Chapter 6, we discussed some of the fundamental properties of the processor-sharing models regardless of scheduling algorithm; specifically we have obtained some of the necessary conditions that a given response function has to follow. But that is only part of the answer; the question as to what are the necessary and sufficient conditions for a given response function to be feasible remains unsolved.

The assumption of zero quantum size is somewhat unrealistic in the real world. The motivation of this assumption is one of simplicity of analysis and in the presentation of results which serve as good approximations to finite-quantum models. Also, the assumption of infinite population may be undesirable if the arrival process somehow depends on the number of customers in the system [45,46]. Clearly, more work needs to be done in these areas as well as the case for more general arrival and service processes so that results that are of more significance could be obtained.

In the past, most of the effort has been spent on the models with single resource. It is true that the central processing unit is probably the most important element in a computer system and the scheduling of its resources is vital to the performance of such a system. And if the size of the memory and the number of I/O devices are adequate then the allocation of the CPU time should be relatively independent of other resources. But in a real computer system, we know this seldom can be attained. The speed of the I/O devices, the size of the main memory, the allocation schemes of the main memory, the size of pages and segments in a paged memory system, program behaviors, as well as the data channels which connect the I/O devices and the memory all have some effect on the performance of a computer system. For example, the effect of a page faulting in a paged memory system may be as important or even more important than the scheduling algorithm in some cases. If a system "thrashes" as described by Denning [47], the central processing unit would be idle most of the time because customers page fault at a very high rate, and it would then make little difference as to what scheduling algorithm is being used in the CPU. Unfortunately, very little work has been done in this area mainly because of the difficulty in analyzing such a queuing system with two or more inter-dependent queuing structures. However, much work must be done in this area in order to get a better understanding of the behavior of the time-sharing computer systems.

Another very important question that has not been answered is the one of optimization where this term itself is yet to be well-defined. So far we have talked about modelling and analyzing of schedulign algorithms that usually all favor the short job over the long one, and in most of the cases the average response time is solved for as the performance measure of the system. We may ask whether this is the only valid criterion for awarding priority in a time-sharing system. If we want to attract customers with long jobs to a computing facility, we must be able to award high priority to them whenever they indicate a willingness to pay the high price. Here we introduce the cost of delay as a performance measure and as a criterion for optimality for time-shared computer systems. Even though the modelling and analysis of a system with different cost functions assigned to different users may become very difficult, a lot of work is needed in this area if some criteria of optimization of models with costs taken from the viewpoint of the system and the users are to be formulated.

BIBLIOGRAPHY

1. McKinney, J.M. "A Survey of Analytical Time-Sharing Models," Computing Surveys 1, No. 2, (June 1969), 105-116.

2. Kleinrock, L. "Analysis of Time-Shared Processor," Nav. Res. Log. Quart. 11, (1964), 59-73.

3. Gibson, C.T. "Time-Sharing in the IBM System/360:Model 67," Proceedings of the Spring Joint Computer Conference, (May 1966), 61-78.

4. Kleinrock, L. "Time-Shared Systems: A Theoretica' Treatment," J.ACM 14, No. 2, (April 1967), 242-261.

5. Coffman, E.G. and Kleinrock, L. "Feedback Queuing Models for Time-Shared Systems," J.ACM 15, No. 4, (October 1968), 549-576.

6. Kleinrock, L. "A Delay Dependent Queue Discipline," Nav. Res. Log. Quart. 2, (1964), 329-341.

7. Brockmeyer, E., Halstrom, H.L., and Jensen, A. "The Life and Work of A.K. Erlang," Translation of the Danish Academy of Technical Sciences, No. 2, (1948), 1-277.

8. Pollaczek, F. "Lieber eine Aufgabe der Wahrscheinlichkeits--Theorie I," Mathematische Zeitschrift 32, 64-100, (1930) ibid II, 32 (1930), 729-750.

9. __________. "Jeber das Warteproblem," Mathematische Zeitschrift 38, (1934), 429-537.

10. Kolmogorov, A.N. "Sur le probleme d'Attente, "Mathematicheskii Sbornik 38, (1931), 101-106.

11. Kendall, D.B. "Some Problems in the Theory of Queues," Journal of the Royal Statistical Society, Series B XIII, No. 2, No. 2, (1951), 151-185.

12. Kendall, D.B. "Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of Imbedded Markov Chains," The Annals of Mathematical Statistics 24, (1953), 338-354.

13. Lindley, D.V. "The Theory of Queues with a Single Server," Proceedings of the Cambridge Philosophical Society 48, (1952), 277-289.

14. Takacs, L. Introduction to the Theory of Queues, Oxford U. Press, New York (1962).

15. Takacs, L. "A Single Server Queue with Feedback," Bell System Technical Journal 42 (March 1963).

16. ________. "Transient Behavior of Single-Server Queuing Processes with Recurrent Input and Exponentially Distributed Service Times," Operations Research 8, No. 1, (January-February 1960), 231-245.

17. ________. "Investigation of Waiting Time Problems by Reduction to Markov Processes," ACTA Mathematica Academiae Scientiarum Hungaricae 6, (1955), 101-129.

18. Little, J.D.C. "A Proof of the Queuing Formula L = λW," Operations Research 9, No. 2, (March-April 1961), 383-387.

19. Cox, D.R., and Smith, W.L. Queues, Methuen & Co., Ltd. (1961).

20. Kleinrock, L. "Swap-Time Consideration in Time-Shared Systems," IEEE Transactions on Computers C-19, No. 6, (June 1970).

21. Schrage, L.E. and Miller, L.W. "The Queue M/G/I with the Shortest Remaining Processing Time Discipline," Operations Research 14, No. 14, (July-August 1966), 670-684.

22. Kleinrock, L. Communication Nets: Stochastic Message Flow and Delay, McGraw-Hill, New York, (1964).

23. Sakata, M., Noguchi, S., and Oizumi, J. "On the Round Robin Procedure for Time-Sharing Systems," The Record of Electrical and Communication Engineering Conversazione, Tohoku University, Japan, 37, No. 2, (1968).

24. ________________________________. "An Analysis of the M/G/1 Queue Under the Round Robin Scheduling," Research Institute of Electrical Communications, Tohoku University, Japan (1969).

25. Coffman. E.G., Muntz, R.R., and Trotter, H. "Waiting Time Distribution for Processor-Sharing Systems," J.ACM 17, No. 1, (January 1970), 123-130.

26. Kleinrock, L. "A Continuum of Time-Sharing Scheduling Algorithms," Proceedings of the Spring Joint Conference (May 1970), 453-458.

27. Schrage, L.E. "The Queue M/G/I with Feedback to Lower Priority Queues," Management Science 13, (1967), 466-471.

28. Shemer, J.E. "Some Mathematical Considerations of Time-Sharing Scheduling Algorithms," J.ACM 14, No. 2, (April 1967), 262-272.

29. Kleinrock, L. and Muntz, R.R. "Multilevel Processor-Sharing Queuing Models for Time-Shared Models," Proc. of the Sixth International Telegraphic Congress, Munich, Germany, (August 1970), 341/1-341/8.

30. Kleinrock, L. and Coffman, E.G. "Distribution of Attained Service in Time-Shared Systems," J. Comput. Syst. Sci. 1, No. 3, (October 1967), 287-298.

31. Scherr, A.L. An Analysis of Time-Shared Computer System, MIT Press, Cambridge, Mass., (1967).

32. Kleinrock, L. "Certain Analytic Results for Time-Shares Processors," Proc. 1968 IFIP Cong., (1968), 838-845.

33. Greenberger, M. "The Priority Problem and Computer Time-Sharing," Management Science 12, (1966), 888-906.

34. Adiri, I. and Avi-Itzhak, B. "A Time-Sharing Queue with a Finite Number of Customers," Oper. Res. and Econ. Mimeo. Series, No. 10, Dep. of Indust. and Mgt. Eng., Technion, Israel Institute of Technology, Haifa, Israel (February 1968).

35. Krishamoorthi, B. and Wood, R.C. "Time-Shared Computer Operations with Both Interarrival and Service Times Exponential," J.ACM 13, No. 3, (July 1966), 317-338.

36. Kleinrock, L. "A Conservation Law for a Wide Class of Queuing Disciplines," Nav. Res. Log. Quart. 12, (1965), 181-192.

37. James, M.S., Smith, G.M., and Wolford, J.C. Analog and Digital Computer Method in Engineering Analysis. International Textbook Company, Scranton, Pennsylvania, (1964).

38. Marshall, K.T. "Bounds for Some Generalizations of the GI/G/I Queue," Operations Research 16, No. 4, (July-August 1968), 841-848.

39. Marshall, K.T. "Some Inequalities in Queuing," Operations Research 16, No. 3, (May-June 1969), 105-116.

40. Kingman, J.F.C. "Some Inequalities for the GI/G/I Queue," Biometrika 49, (1968), 315-324.

41. Iglehart, D.L. "Diffusion Approximations in Applied Probability," Mathematics of the Decision Sciences (Part 2), G.B. Dentzig and A.F. Veinott, Jr., Editors; American Math. Soc., Providence, R.I., (1968).

42. Daley, D.J. and Moran, P.A.P. "Two-Sided Inequalities for Waiting Time and Queue Size Distributions in GI/G/1," Theory of Probability XIII, No. 2, (1968), 338-341.

43. Gaver, D.P. "Diffusion Approximations and Models for Certain Congestion Problems," Journal of Applied Probability 5, (1968), 607-623.

44. Kleinrock, L., Muntz, R.R., and Hsu, J. "Tight Bounds on the Average Response Time for Time-Shared Computer Systems," Proceedings of the IFIP Congress, (1971).

45. Buzen, J. "Analysis of System Bottlenecks Using a Queuing Network Model," ACM SIGOPS Workshop on System Performance Evaluation, (April 1971), 82-103.

46. Gordon, W.J. and Newell, G.F. "Closed Queuing Systems with Exponential Service," Operations Research 15, No. 2, (March-April 1967), 254-265.

47. Denning, P.J. "Thrashing: Its Causes and Prevention," Proceedings of the Fall Joint Computer Conference, (1968), 915-922.

APPENDIX A

THEOREMS AND PROOFS FOR THEOREMS IN CHAPTER 3

A.1 Theorem 3.2 and Its Proof

Theorem 3.2 For any customer requiring t seconds of service, the time he spends in the queue box is independent of the time he spends in the service box (or independent of the time he wasted in the service box because t is not a random variable).

Proof: We prove this theorem by using an argument for busy period distributions [19]. The Laplace transform $P^*(s)$ of the distribution of the busy periods for the M/G/1 system is given as (see Eq. (2.12))

$$P^*(s) = B^* [s + \lambda - \lambda P^*(s)] \tag{A.1}$$

with mean value g_1 and second moment g_2 as follows:

$$g_1 = \frac{1/\mu}{1 - \rho} \tag{A.2}$$

$$g_2 = \frac{\overline{t^2}}{(1 - \rho)^3} \tag{A.3}$$

where

$$\overline{t^2} = \lim_{s \to 0} \frac{\partial^2 B^*(s)}{\partial s^2} = \text{second moment of service time distribution} \tag{A.4}$$

Since all the work has to be done in the service box and since the arrival process to the service box when it is not idle is Poisson, the service box itself can be regarded as an M/G/1 system with average arrival rate λ'. The Laplace transform $P^*(s)$ of the busy period distribution can also be expressed as

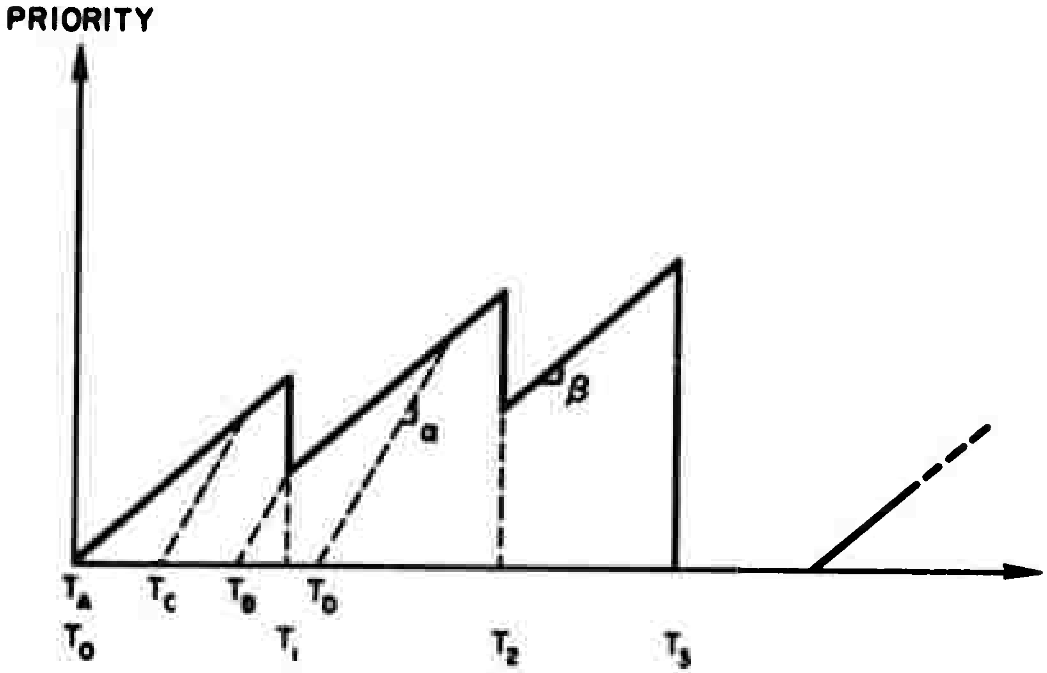
$$P^*(s) = B^*[s + \lambda' - \lambda' P^*(s)] \tag{A.5}$$

with mean value g_1' and second moment g_2'

Preceding page blank

$$g_1' = \frac{1/\mu}{1 - \rho'} \tag{A.6}$$

$$g_2' = \frac{\overline{t^2}}{(1 - \rho')^3} \tag{A.7}$$



BUSY PERIOD DISTRIBUTION OF AN SSA SYSTEM

FIGURE A-1

Refer to Figure A-1; a customer arrives to an empty system at time T_0 and starts a busy period in the system as well as in the service box because he does not have to spend any time in the queue box. From T_0 to T_1, the service box is busy as more customers arrive to the system. Suppose that the last customer in the service box leaves at T_1; that marks the end of this "small" busy period. If the queue box is not empty at this time, the customer with the highest priority will be admitted immediately into the service box, thus starting another small busy period. The priority of the service box will adjust to whatever priority this customer has; usually this means a drop of priority as shown by Figure A-1. If, when the last customer in the service box departs and there is no customer waiting in the queue box, the system goes idle and this means the end of a busy period of the system. In order to differentiate this with the small busy period we mentioned earlier, we call this a <u>large</u> busy period. In short, a large busy period is the time interval when the system is busy; and a small busy period is the time interval the service box is busy (with no drop in priorities). Obviously, a large busy period usually contains one or more small busy periods. For an M/G/1 system with average arrival rate λ and service rate μ , the average length of a large busy period is $\frac{1/\mu}{1 - \rho}$; similarly, the average length of a small busy period is $\frac{1/\mu}{1 - \rho'}$. Therefore, on the average, there are $(\frac{1 - \rho'}{1 - \rho})$ small busy periods in a large busy period. In Figure A-1, the interval (T_0,T_3) is a large busy period, while time intervals (T_0,T_1), (T_1,T_2), and (T_2,T_3) are small busy periods.

In order to prove the theorem, let us refer to Figure A-1. Assume that customer A starts a large as well as a small busy period; customer B enters the service box at T_1 and, therefore, starts a small busy period but not a large one. Thus, customer A does not have to wait in the queue box while customer B does. After they enter the service box (at different times), customers A and B will see the same environment (M/G/1 system with average arrival rate λ'). There is no way to differentiate these two customers statistically from the time they enter the service box because they both start a small busy period and all small busy periods are identically distributed, as expressed by

Eq. (A.5) [19]. Thus, for customers A and B, the time they spend in the service box must be independent of the time they spend in the queue box.

Next, let us look at customers C and D under the assumption the distances $(T_C - T_0)$ and $(T_D - T_B)$ are the same, that is, C and D enter their respective small busy periods at the same corresponding time (i.e., the service box has been busy for the same amount of time since the start of current small busy period). Then, as far as the time spent in the service box is concerned, there is no difference between C and D statistically because the small busy periods are identically distributed, but their waiting times in the queue box are different (as represented by their respective priorities when they enter the service box) as depicted by Figure A-1. Therefore, for customers C and D, the theorem holds true; but C and D can be customers, so the independent assumption must be true for every customer.

Q.E.D.

A.2 Theorem 3.3 and Its Proof

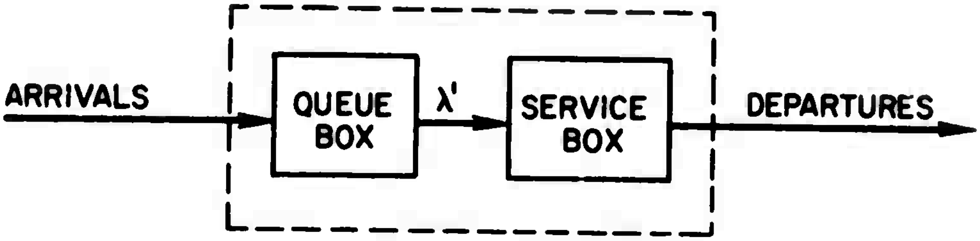
Theorem 3.3 The Laplace transform Q*(s) of the density function of the waiting time spent in the queue box by a customer requiring t seconds of service time (actually, it is independent of t as we explained earlier) is

$$Q^*(s) = \frac{(1 - \rho)}{(1 - \rho')} \cdot \frac{\lambda' B^*(s) - \lambda' + s}{\lambda B^*(s) - \lambda + s}$$

with first moment equal to

$$W_q = \lim_{s \to 0} - \frac{\partial Q^*(s)}{\partial s} = \frac{\lambda \overline{t^2}}{2(1 - \rho)} - \frac{\lambda' \overline{t^2}}{2(1 - \rho')}$$

Proof:



DECOMPOSITION OF THE SSA SYSTEM

FIGURE A-2

As we said earlier, the scheduling algorithm in the service box will not affect the waiting time distribution as long as no feedback from the service box to the queue box is possible. For convenience, let us assume that an FCFS scheduling algorithm is being used, namely, after a customer enters the service box, he will be served on an FCFS basis to completion. This makes the whole system FCFS. The Laplace transform, S*(t,s), of the equilibrium response time distribution for an FCFS system is well known [19], namely

$$S^*(t,s) = B^*(s) \frac{(1 - \rho)s}{B^*(s) - \lambda + s} \qquad (A.8)$$

After the tagged customer enters the service box, he is in another FCFS system with the average arrival rate λ' instead of λ, thus we can easily get $Y^*(t,s)$ as

$$Y^*(t,s) = B^*(s) \frac{(1 - \rho')s}{\lambda' B^*(s) - \lambda' + s} \qquad (A.9)$$

From the independent property proved in Theorem 3.2, $Q^*(s)$ can easily be obtained as the ratio of $S^*(t,s)$ and $Y^*(t,s)$

$$Q^*(s) = \frac{S^*(t,s)}{Y^*(t,s)}$$

$$= \frac{(1 - \rho)}{(1 - \rho')} \; \frac{\lambda' B^*(s) - \lambda' + s}{\lambda\, B^*(s) - \lambda + s} \qquad \text{(A.10)}$$

Q.E.D.

APPENDIX B

THEOREM 4.1 AND ITS PROOF

Theorem 4.1 and Its Proof

Theorem 4.1 The average response time $T(t)$ for the system with the scheduling algorithm defined by $g(t) = ge^{-gt}$, is the solution of the following integro-differential equation

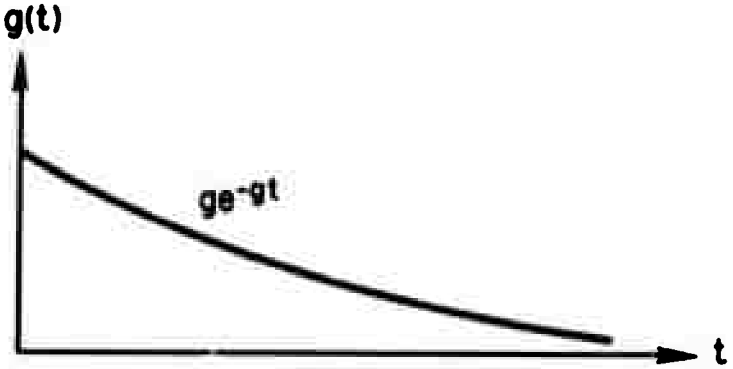
$$(1-\rho)T(t) = t + \frac{\rho/\mu}{1-\rho} - \rho\int_0^\infty T'(t)\left[e^{g\tau} + e^{gt} - 1\right]^{-\mu/g} d\tau$$

$$- \rho\int_0^t T'(\tau)\left[e^{gt} - e^{g\tau} + 1\right]^{-\mu/g} d\tau$$

Proof: In Section 2.2.9 we discussed the attained service and remarked that the density of customers having obtained t seconds of attained service is given by [30]

$$n(t) = \lambda T'(t)\,[1 - B(t)] \qquad \text{(B.1)}$$

Although $n(t)$ is not available to us, nevertheless, we can use it as an intermediate step for the calculation of $T(t)$.



QUANTUM ASSIGNMENT AS FUNCTION OF ATTAINED SERVICE TIME

FIGURE B-1

In Figure B-1 we plot $g(t)$ against t, where $g(t)$ denotes the relative rate of attaining service for customers with different amounts of attained service times. What we are interested in is that given an elapsed time interval T, what are the relative amounts of services attained by customers who have different accumulated amounts of service times to begin with. In other words, we wish to consider the time interval during which a customer (let us call him customer A) remains in the system and gains t seconds of service. During this interval, we are interested in how much service time x a customer (let us call him customer B) can get during this same time interval given that customer B has received τ seconds of service just prior to the beginning of this time interval. In order to calculate x, let us first make the following observation:

$$\frac{dy}{d\tilde{T}(y)} = \tilde{f}(y) \tag{B.2}$$

$$\frac{\dfrac{dy}{d\tilde{T}(y)}}{e^{-gy}} = \frac{\dfrac{dz}{d\tilde{T}(z)}}{e^{-gz}} = c \tag{B.3}$$

where x and y are arbitrary time instants on the time axis and $\tilde{T}(y)$ is a random variable denoting the response time for getting y seconds of service time. Equation (B.2) states that the rate of gaining service for a customer with y seconds of attained service is equal to $\tilde{f}(y)$, the fraction of the total service facility that is allocated to him. Equation (B.3) says that a customer with y seconds of service gains service at a rate proportional to e^{-gy}; and a customer with z seconds of attained service at a rate proportional to e^{-gz}. The proportionally constant is given as c which does not depend on either y or z. Equation (B.3), of course, just restates the definition of the algorithm as defined by Eq. (4.1) and is a direct consequence of Eqs. (4.2) and

(B.2). Since the time interval during which customer A gains t seconds of service is the same as that during which customer B gains x seconds of service, we get the following equation

$$\int_0^t \frac{d\tilde{T}(y)}{dy}\,dy = \int_\tau^{\tau+x} \frac{d\tilde{T}(z)}{dz}\,dz \tag{B.4}$$

Substituting Eq. (B.3) into Eq. (B.4), we get

$$c\int_0^t e^{gy}dy = c\int_\tau^{\tau+x} e^{gz}dz \tag{B.5}$$

or

$$e^{gt} - 1 = e^{g(\tau+x)} - e^{g\tau} \tag{B.6}$$

Rearranging Eq. (B.6), we get

$$x = \frac{1}{g}\ln[e^{-g\tau}(e^{gt} - 1) + 1] \tag{B.7}$$

Thus, in order for a tagged customer to get t seconds of service time, the service facility has to serve customers who are already in the system with τ seconds of attained service when the tagged customer arrives to the system each with up to x seconds of service. Since the density of customers in the system is assumed to be distributed as $n(\tau)$, the total amount of work $U(t)$ which needs to be done to them is calculated as

$$U(t) = \int_0^\infty n(\tau)\,[\bar{t}_{<x}]\,d\tau$$

$$= \int_0^\infty n(\tau)\,(1/\mu)\,[1 - e^{-\mu x}]\,d\tau$$

$$= \int_0^{\infty} n(\tau)(1/\mu)\left| 1 - [e^{-g\tau}(e^{gt} - 1) + 1]^{-\mu/g} \right| d\tau \qquad \text{(B.8)}$$

where $\bar{t}_x$ is the average amount of service time with the service distribution truncated at x.

$U(t)$ takes care of all the customers who are in the system prior to the arrival of the tagged customer. The next step for us is to calculate the amount of work $V(t)$ that needs to be done to those customers who arrive to the system later than the tagged customer but before he leaves the system with his t seconds of service. Let us assume that one of those later comers (let us call him customer C) arrives to the system when the tagged customer (customer A) has accumulated exactly τ seconds of service. We wish to find out how much service w that customer C is going to get during the same time interval that customer A gets served by $(t-\tau)$ seconds. Eq. (B.4) can be readily modified as

$$\int_0^{w} \frac{d\tilde{T}(y)}{dy}\, dy = \int_{\tau}^{t} \frac{d\tilde{T}(z)}{dz}\, dz \qquad \text{(B.9)}$$

and Eq. (B.5) becomes

$$\int_0^{w} e^{gy} dy = \int_{\tau}^{t} e^{gz} dz$$

or

$$e^{gw} - 1 = e^{gt} - e^{g\tau} \qquad \text{(B.10)}$$

Rearranging Eq. (B.10), w can be calculated as

$$W = \frac{1}{g}\ln[e^{gt} - e^{g\tau} + 1] \qquad \text{(B.11)}$$

w is the maximum amount of service customer C can get during the interval when the tagged customer accumulates his attained service time from τ to t seconds. The actual service time customer C gets before he leaves the system is, on the average, smaller than w. It can easily be calculated for exponentially distributed service times as

$$\bar{t}_{<w} = \frac{1}{\mu}(1 - e^{-\mu w}) \tag{B.12}$$

During the differential time interval when the tagged customer gets from τ to $\tau + d\tau$ seconds of attained service, there are, on the average, $T'(\tau)d\tau$ (Little's Result) new arrivals coming to the system, therefore, $V(t)$ can be expressed in the following equation:

$$V(t) = \int_0^t \lambda T'(\tau) [\bar{t}_{<w}] d\tau$$

$$= \int_0^t \lambda T'(\tau) \frac{1}{\mu} [1 - e^{-\mu w}] d\tau$$

$$= \int_0^t \rho T'(\tau) \left| 1 - [e^{gt} - e^{g\tau} + 1]^{-\mu/g} \right| d\tau \tag{B.13}$$

The average response time $T(t)$ (i.e., the total average time the tagged customer spends in the system in order to get served for t seconds) is the sum of $t, U(t)$ and $V(t)$. From Eq. (B.7) and (B.13), $T(t)$ is given by

$$T(t) = T + A(t) + B(t)$$

$$= t + \int_0^\infty n(\tau) \frac{1}{\mu} \left| 1 - [e^{-g\tau}(e^{gt} - 1) + 1]^{-\mu/g} \right| d\tau$$

$$+ \rho \int_0^t T'(\tau) \left| 1 - [e^{gt} - e^{g\tau} + 1]^{-\mu/g} \right| d\tau \qquad \text{(B.14)}$$

$n(\tau)$ is given by Eq. (B.1), substituting $n(\tau)$ into Eq. (B.14) and simplifying, we get

$$T(t) = t + \frac{1}{\mu} \int_0^\infty n(\tau) d\tau - \rho \int_0^\infty T'(\tau) e^{-\mu\tau} [e^{-g\tau}(e^{gt} - 1) + 1]^{-\mu/g} d\tau$$

$$+ \rho T(t) - \rho \int_0^t T'(\tau) [e^{gt} - e^{g\tau} + 1]^{-\mu/g} d\tau \qquad \text{(B.15)}$$

For M/M/1 systems (from Eq. (2.5)) the average number in the system is given by

$$\int_0^\infty n(\tau) d\tau = \frac{\rho}{1 - \rho} \qquad \text{(B.16)}$$

Equation (B.15) becomes

$$T(t)(1 - \rho) = t + \frac{\rho/\mu}{1 - \rho} - \rho \int_0^\infty T'(\tau) e^{-\mu\tau} e^{\mu\tau} [e^{gt} - 1 + e^{g\tau}]^{-\mu/g} d\tau$$

$$- \rho \int_0^t T'(\tau) [e^{gt} - e^{g\tau} + 1]^{-\mu/g} d\tau \qquad \text{(B.17)}$$

Finally, we get

$$T(t)(1 - \rho) = t + \frac{\rho/\mu}{1 - \rho} - \rho \int_0^\infty T'(\tau) [e^{gt} + e^{g\tau} - 1]^{-\mu/g} d\tau$$

$$- \rho \int_0^t T'(\tau) [e^{gt} - e^{g\tau} + 1]^{-\mu/g} d\tau \qquad \text{(B.18)}$$

Q.E.D.

APPENDIX C

THEOREMS AND PROOFS FOR THEOREMS IN CHAPTER 6

C.1 Theorem 6.1 and Its Proof

Theorem 6.1 $W(t)$ is a nondecreasing function of t or equivalently

$$W'(t) \equiv \frac{dW(t)}{dt} \geq 0$$

Proof: We are considering scheduling disciplines in which each job is characterized by (1) its attained service time, t_s and (2) its elapsed waiting time, t_w. Therefore, the state of the system is the number of jobs in the system and t_s and t_w for each job. A particular schedusing discipline may effectively ignore one or both of these parameters, but this information is assumed to be available for each job. Because scheduling decisions are made only on the basis of these two parameters, the following statement is self-evident. The history of a job requiring $t_1 \geq t$ seconds of service from the time of its arrival at the system until it has received t seconds of service is independent of the exact value of t_1. A direct consequence of this fact is that $W(t)$ is a nondecreasing function or equivalently

$$W'(t) \equiv \frac{dW(t)}{dt} \geq 0 \qquad \text{(C.1)}$$

Q.E.D.

C.2 Theorem 6.2 and Its Proof

Theorem 6.2 There is a conservation law that $W(t)$ has to satisfy, namely

$$\int_0^\infty W(t)\,[1 - B(t)]\,dt = \frac{\rho\overline{t^2}}{2(1-\rho)}$$

For $T(t)$, the conservation law becomes

$$\int_0^\infty T(t)\,[1 - B(t)]\,dt = \frac{\overline{t^2}}{2(1 - \rho)}$$

<u>Proof</u>: From [30] we have that

$$n(t) = \lambda[1 - B(t)]\,[W'(t) + 1] \qquad (C.2)$$

where $n(t)$ is the density of jobs in the system with t seconds of attained service time. We define the "work" in the system at time t as the additional time required to empty the system if no new arrivals are permitted entry; this is also referred to as the "unfinished work" and as the"virtual waiting time." The mean work $\overline{W}$ in the system can be expressed as

$$\overline{W} = \int_0^\infty n(t)\,E[\text{remaining service time for a job with attained service time of } t]\,dt$$

or

$$\overline{W} = \int_0^\infty n(t) \int_t^\infty (\tau - t)\,\frac{dB(\tau)}{1 - B(t)}\,dt$$

Substituting from (C.2)

$$\overline{W} = \lambda \int_0^\infty (\overline{W}'(t) + 1) \int_t^\infty (\tau - t)\,dB(\tau)\,dt$$

By changing the order of integration

$$\overline{W} = \lambda \int_0^\infty \left[\int_0^\tau (W'(t) + 1)(\tau - t)\,dt \right] dB(\tau) \qquad (C.3)$$

Integrating the inner integral by parts

$$\int_0^\tau (W'(t) + 1)(\tau - t)\,dt = (\tau - t)(W(t) + t)\Big|_0^\tau + \int_0^\tau [W(t) + t]\,dt$$

$$= \int_0^\tau [W(t) + t]\,dt$$

Substituting into Eq. (C.3)

$$\overline{W} = \lambda \int_0^\infty \int_0^\tau [W(t) + t] dt \, dB(\tau)$$

Again changing the order of integration

$$\overline{W} = \lambda \int_0^\infty [W(t) + t] \int_t^\infty dB(\tau) dt$$

But integrating by parts, we have that

$$\int_0^\infty t[1 - B(t)] dt = \frac{\overline{t^2}}{2}$$

Moreover, the mean work in the system is known

$$\overline{W} = \frac{\lambda \overline{t^2}}{2(1 - \rho)} \qquad \text{(C.4)}$$

Thus we have the following <u>conservation laws</u> for $T(t)$ and $W(t)$:

$$\frac{\overline{t^2}}{2(1 - \rho)} = \int_0^\infty T(t) [1 - B(t)] dt \qquad \text{(C.5)}$$

and

$$\frac{\rho \overline{t^2}}{2(1 - \rho)} = \int_0^\infty W(t) [1 - B(t)] dt \qquad \text{(C.6)}$$

Q.E.D.

C.3 Theorem 6.3 and Its Proof

Theorem 6.3 The lower bound $W_\ell(t)$ of waiting time functions is given by the waiting time for the FCFS discipline with the service time distribution truncated at t, namely

$$W_\ell(t) = \frac{\lambda \overline{t^2}_{<t}}{2(1 - \rho_{<t})}$$

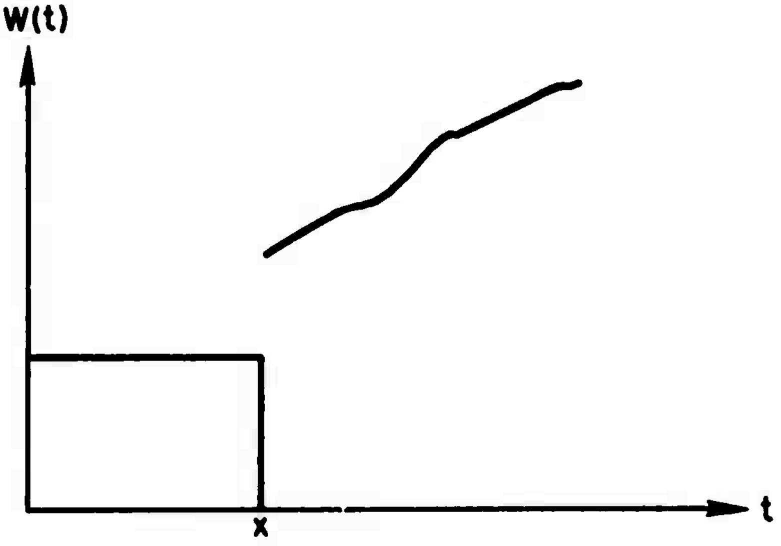
Note that $W_\ell(0) = 0$ and that $W_\ell(\infty) = W_{FCFS}$ (the average waiting time for the FCFS system as given in Section 2.2.1), also $W'_\ell(0) = W'_\ell(\infty) = 0$.

Proof: We claim that to minimize $W(x)$ the scheduling discipline must

1. never service jobs with attained service time greater than or equal to x while there are jobs in the system with attained service time less than x, and

2. never preempt a job once it has been selected for service until it has at least x seconds of attained service time,

Under these conditions the response function in the interval $(0,x)$ is just the response function for a nonpreemptive system with service times truncated at x. For convenience we will assume a FCFS scheduling discipline. In this case the response function (denoted $W_{FCFS-x}(t)$ has the form shown in Figure C-1 (see Section 2.2.9). Note that $W'_{FCFS-x}(t) = 0$ over $(0,x)$. The scheduling of jobs with attained service time greater than x is of no concern in this argument as long as condition 1 is maintained.



RESPONSE FOR FCFS UP TO X SECONDS OF SERVICE

FIGURE C-1

Let $\bar{W}_x$ be the mean work in the system excluding work to be done on jobs beyond providing x seconds of attained service to each. In other words, if a job requires $t > x$ seconds of service and has received $y < x$ seconds of service, its contribution to $\bar{W}_x$ is $x - y$. By the same method used to derive Eq. (C.5) it can be shown that

$$\bar{W}_x = \lambda\int_0^x [W(t) + t][1 - B(t)]dt$$

Now since $W_{FCFS-x}(t)$ has minimum slope (i.e., 0) only over the interval $(0,x)$, and due to the monotonicity given in Eq. (C.1), if any other response curve $W(t)$ is such that $W(x) < W_{FCFS-x}(x)$ it must be such that $W(t) < W_{FCFS-x}(t)$ for $0 \leq t < x$. But under condition 1 above, $\bar{W}_x$ has its minimum value since work in this class is continuously decreased at maximum rate whenever there is such work in the system. Therefore, for any $W(t)$,

$$\lambda\int_0^x [W(t) + t][1 - B(t)dt$$
$$\geq \lambda\int_0^x [W_{FCFS-x}(t) + t][1 - B(t)]dt$$

Thus we conclude that $W(t) < W_{FCFS-x}(t)$ in $(0,x)$ is impossible and therefore $W(x) \geq W_{FCFS-x}(x)$.

The lower bound $W_\ell(t)$ is given by the waiting time for the FCFS discipline with the service times truncated at t, namely

$$W_\ell(t) = \frac{\lambda \overline{t^2_{<t}}}{2(1 - \rho_{<t})} \tag{C.7}$$

Q.E.D.

C.4 Theorem 6.4 and Its Proof

Theorem 6.4 The upper bound $W_u(t)$ of waiting time functions is given as

$$W_u(t) = \frac{\lambda\overline{t^2}}{2(1-\rho_{<t})} + \frac{t\rho_{<t}}{1-\rho_{<t}}$$

One scheduling discipline which gives $W_u(t)$ is a two-level system with both levels served FCFS and switching point (see Chapter 6) at t. Note that $W_u(0) = W_\ell(\infty) = W_{FCFS}$, that $W'_u(0) = 0$ and that $W'_u(\infty) = \frac{\rho}{1-\rho}$.

Proof: In this case we begin with a discrete time system.

Assume that the service time distribution is of the form

$$\Pr[\text{service time} = kq] = p_k \qquad k = 1,2,3,\ldots$$

where q is the quantum size. Therefore, the only possible service time requirements are multiples of q. We shall also assume that arrivals may take place only during the instant before the end of a quantum and that the processor is assigned to a job for a quantum at a time. The probability that an arrival takes place at the end of a quantum is λq so that the mean arrival rate is λ. It should be clear that any continuous service time distribution can be approximated arbitrarily closely by a discrete time distrubution by letting q approach 0. Also, these restrictions on the service discipline and arrival mechanism are effectively eliminated when $q \to 0$. In this discrete time model our goal is to maximize $W(kq)$.

We claim that the following scheduling rule is necessary and sufficient to maximize $W(kq)$: no allocation of a kth quantum is made to any job when there is some other job in the system waiting for its

its jth quantum where $j \neq k$. We note in passing that many scheduling disciplines will satisfy this rule.

We relabel the time axis so that $t = 0$ at an arbitrary point in some idle period. The times at which some job is allocated to a kth quantum we call "critical times." Let c_i be the time that the i^{th} critical time occurs. We wish to maximize $\bar{c}_\ell$ (the average of c_ℓ) for some fixed ℓ, and we will show that to accomplish this it is necessary and sufficient to satisfy the condition that at the ℓ^{th} critical time no job is waiting for a j^{th} quantum where $j \neq k$. Certainly this condition is necessary since if a proposed scheduling discipline did not have this property then c_ℓ can easily be increased when the condition is not satisfied as follows: follow the proposed schedule until the point where the ℓ^{th} critical time would occur and then assign a quantum to a job waiting for its j^{th} $(\neq k)$ quantum.

Since we have already shown necessity, to prove the sufficiency of the condition for maximizing $\bar{c}_\ell$, we need only show that any schedule satisfying the condition yields the same value for $\bar{c}_\ell$. Let A be any scheduling algorithm which satisfies the rule that at the ℓ^{th} critical time no job is waiting for a j^{th} quantum where $j \neq k$. Let a_ℓ be the time at which the ℓ^{th} job arrives which will require at least kq seconds of service. The state of the system at a_ℓ will, in general, depend on the algorithm A. In particular, the number of critical times that have occurred prior to a_ℓ (let this be s) is a function of A. Let $E_A[c_\ell - a_\ell |$ state of systems at $a_\ell]$ be the expected value of $c_\ell - a_\ell$ under algorithm A conditioned on the state of the system at a_ℓ. The state of the system is given by the number of jobs in the system, the attained service time of each job in the system and s, the number of critical times that have occurred. Thus we have

$$E_A[c_\ell - a_\ell \mid \text{state of system at } a_\ell]$$

$$= [\text{remaining work in system not requiring a } k^{th} \text{ quantum} \mid \text{state of system at } a_\ell]$$

$$+ (\ell - s - 1)E[\text{remaining service time for job with } (k-1)q \text{ seconds of attained service}]$$

$$+ (k - 1)q$$

$$+ \lambda\bar{t}_{<(k-1)q}E_A[c_\ell - a_\ell \mid \text{state of the system at } a_\ell] \qquad (C.8)$$

But the sum of the first two terms on the righthand side of this equation is equal to the expected amount of work in the system at a_ℓ given the state at a_ℓ. Thus

$$E_A[c_\ell - a_\ell \mid \text{state of system at } a_\ell]$$

$$= E_A[\text{work in system at } a_\ell \text{ state at } a_\ell]$$

$$+ (k - 1)q$$

$$+ \lambda\bar{t}_{<(k-1)q}E_A[C_\ell - a_\ell \mid \text{state of system at } a_\ell]$$

Removing the condition on the state of the system at a_ℓ we have

$$E_A[c_\ell - a_\ell] = E_A[\text{work in the system at } a_\ell]$$

$$+ (k - 1)q + \lambda\bar{t}_{<(k-1)q}E_A[c_\ell - a_\ell]$$

or

$$E_A[c_\ell - a_\ell] = \frac{E_A[\text{work in system at } a_\ell] + (k-1)q}{1 - \lambda\bar{t}_{<(k-1)q}}$$

But E_A[work in system at a_ℓ] is not a function of the particular scheduling algorithm and therefore $E_A[c_\ell - a_\ell]$ does not depend on A. Since $E[c_\ell] = E[c \ - a_\ell] + E[a_\ell]$ and the right-hand side is independent of A, $E[c_\ell]$ is independent of A. Note that the form of Eq. (C.8) depended on A having the property that at c_ℓ there are no jobs in the system waiting for a j^{th} quantum where $j \neq k$. We have now shown that this condition is necessary and sufficient to maximize $E[c_\ell](= \bar{c}_\ell)$.

We now show that the general scheduling rule to maximize W(kq) is the same rule which maximizes $\bar{c}_\ell$ applied for all ℓ. We have

$$W(kq) = \lim_{n\to\infty} \frac{\sum_{\ell=1}^{n} \bar{c}_\ell - \sum_{\ell=1}^{n} \bar{a}_\ell}{n} \tag{C.9}$$

The $\bar{a}_\ell$ are independent of the scheduling discipline and the proposed scheduling rule is necessary and sufficient to individually maximize the $\bar{c}_\ell$. Therefore, the same rule is necessary and sufficient to maximize W(kq), which establishes our earlier claim.

It should be clear that in a continuous time system we can approach the maximum of W(x) by the following rule: no job with attained service time in the open interval $(x - \varepsilon, x)$ (for $\varepsilon > 0$) is serviced while there is a job waiting for service which has attained service time outside this interval. By permitting ε to shrink to zero, we approach the maximum for W(x).

One scheduling discipline which maximizes W(x) is the two-level system in which jobs are served FCFS in the first level up to x^- seconds of attained service. A job which does not finish is placed in the second level queue. The second queue is serviced FCFS to completion. The second queue has a lower priority and is only serviced when the first queue is empty (see the ML systems described in Section 3). This queueing system satisfies the condition for maximizing $W(x)$ and therefore from Eq. (2.57) we have

$$W_u(t) = \frac{\lambda\overline{t^2}}{2(1 - \rho_{<t})(1 - \rho)} + \frac{t\rho_{<t}}{1 - \rho_{<t}} \tag{C.10}$$

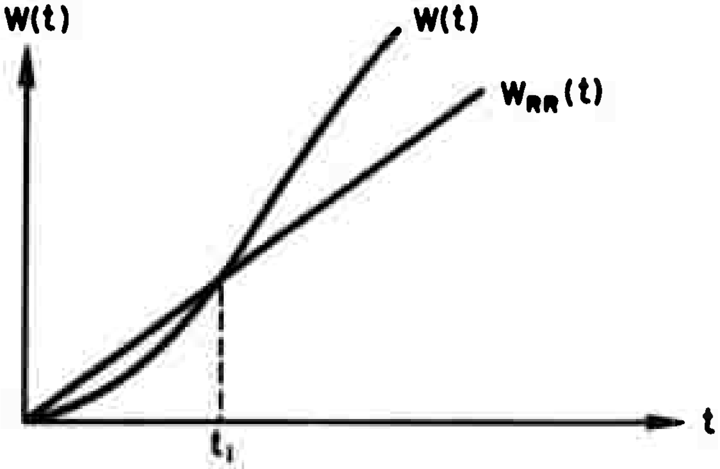
Q.E.D.

C.5 Theorem 6.5 and Its Proof

Theorem 6.5 For a response time function $W(t)$ which is continuously differentiable, $W'(t) = \frac{dW(t)}{dt}$ cannot be monotonically non-decreasing with t.

Proof: Let us prove this theorem by contradiction. Suppose that we can find a $W'(t)$ which is monotonically non-decreasing with t, then $W(t)$ either does not intersect $W_{RR}(t)$ which represents the waiting time function for the RR system; or it intersects $W_{RR}(t)$ at only one point as shown in Figure C-2. If $W(t)$ does not intersect with $W_{RR}(t)$ then it either lies entirely above $W(t)$ or entirely below $W(t)$, neither of these two situations is possible because they violate the conservation law as depicted by Eq. (C.5).

On the other hand, if $W(t)$ crosses $W_{RR}(t)$ at t_i, since $W(t)$ is continuously differentiable and non-decreasing, $W'(t)$ has to be larger than $W'_{RR}(t)$ for all $t > t_i$. But in Theorem 6.4, we prove that $W'_u(\infty) = W'_{RR}(t) = \frac{\rho}{1 - \rho}$, thus we have $W'(t) > W'_u(\infty)$ for all $t > t_i$. It means that $W(t)$ increases at a faster rate than the upper bound for all waiting time functions. Sooner or later, $W(t)$ will intersect $W_u(t)$ and then assumes larger value than $W_u(t)$. This, of course, violates the definition of the upper bound.



WAITING TIME FUNCTION POSSIBILITY

FIGURE C-2

Q.E.D.